The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs *

David E. Broockman[†] Joshua L. Kalla[‡] Jasjeet S. Sekhon[§]

May 15, 2017

Abstract

There is increasing interest in experiments where outcomes are measured by surveys and treatments are delivered by a separate mechanism in the real world, such as by mailers, door-to-door canvasses, phone calls, or online ads. However, common designs for such experiments are often prohibitively expensive, vulnerable to bias, and raise ethical concerns. We show how four methodological practices currently uncommon in such experiments have previously undocumented complementarities that can dramatically relax these constraints when at least two are used in combination: 1) online surveys recruited from a defined sampling frame 2) with at least one baseline wave prior to treatment 3) with multiple items combined into an index to measure outcomes and, 4) when possible, a placebo control. We provide a general and extensible framework that allows researchers to determine the most efficient mix of these practices in diverse applications. Two studies then examine how these practices perform empirically. First, we examine the representativeness of online panel respondents recruited from a defined sampling frame and find that their representativeness compares favorably to phone panel respondents. Second, an original experiment successfully implements all four practices in the context of a door-to-door canvassing experiment. We conclude discussing potential extensions.

---

*This paper previously circulated under the title "Testing Theories of Attitude Change With Online Panel Field Experiments." Software for planning an experiment using all four practices we describe is available at `http://experiments.berkeley.edu`. Replication data is available as Broockman, Kalla and Sekhon (2017), at `http://dx.doi.org/10.7910/DVN/EEP5MT`. This work was supported by the NARAL Pro-Choice America Foundation, the Signatures Innovations Fellows program at UC Berkeley, UC Berkeley's Institute for Governmental Studies for supporting this research, and the Office of Naval Research [N00014-15-1-2367]. The studies reported herein were approved by Committee for the Protection of Human Subjects. We thank participants at the 2015 POLMETH meeting and at the University of California, Berkeley's Research Workshop in American Politics for helpful feedback. Additional feedback was provided by Peter Aronow, Rebecca Barter, Kevin Collins, Alex Coppock, Jamie Druckman, Thad Dunning, Donald Green, Christian Fong, Seth Hill, Dan Hopkins, Gabe Lenz, Winston Lin, Chris Mann, David Nickerson, Kellie Ottoboni, Kevin Quinn, Fredrik Sävje, Yotam Shev-Tom, Bradley Spahn, and Laura Stoker. All remaining errors are our own.

[†]Assistant Professor, Stanford Graduate School of Business. `dbroockman@stanford.edu`, `https://people.stanford.edu/dbroock/`.

[‡]Graduate Student, Department of Political Science, University of California, Berkeley. `kalla@berkeley.edu`, `http://polisci.berkeley.edu/people/person/joshua-kalla`.

[§]Robson Professor of Political Science and Statistics, University of California, Berkeley. `sekhon@berkeley.edu`, `http://sekhon.berkeley.edu`.

1 Introduction

Researchers of political psychology, intergroup prejudice, media effects, learning, public health, and more frequently test how randomized stimuli affect outcomes measured in surveys. For example, experiments that measure the effects of randomized stimuli presented in a survey on individuals' responses to questions in the same survey ('survey experiments') constitute a dominant paradigm in political science (Druckman et al. 2006; Sniderman and Grob 1996); leading political science journals publish dozens of survey experiments each year (Hainmueller, Hopkins and Yamamoto 2014, p. 1).

Along with political scientists' strong interests in survey research and experiments, there has been increasing interest in *field* experiments with survey outcomes: experiments where outcomes are measured by surveys but randomized stimuli are delivered by a separate mechanism in the real world, such as by mailers, door-to-door canvasses, phone calls, or online ads. Unfortunately, low rates of survey response and treatment compliance in many countries mean common designs for such experiments present scholars several barriers. First, common designs for such experiments are often infeasibly expensive. For example, an experiment well-powered to detect a 'small' treatment effect of 0.1 standard deviations on a survey outcome could cost under $500 as a survey experiment using Mechanical Turk but easily over $1,000,000 as a field experiment using designs common today (see next section). In addition, the results of such experiments are vulnerable to bias from differential attrition, which occurs when treatments influence survey completion. This has been shown to occur and produce meaningfully large bias, yet is often undetectable with common designs (Bailey, Hopkins and Rogers 2016). Finally, to be well-powered they require real-world intervention on a grand scale, raising ethical concern (Michelson 2016).

This paper makes three related contributions that can help researchers conduct field experiments with survey outcomes that are significantly more feasible, precise, robust, and ethical.

Our first contribution is to describe and analytically decompose previously undocumented com-

plementarities between four methodological practices currently uncommon in such experiments. These are: 1) surveys administered online to a sample recruited from an ex ante defined sampling frame (e.g., Barber et al. 2014), 2) with at least one baseline wave prior to treatment (Iyengar and Vavreck 2012) 3) with multiple measures of outcomes gathered and combined into an index at each wave (Ansolabehere, Rodden and Snyder 2008) and, if possible, 4) a placebo wherein control subjects are contacted with an unrelated appeal (Nickerson 2005*b*).

The complementarities between these four practices have not been previously documented to our knowledge yet can yield extremely large gains. These practices are not novel on their own. Moreover, in common cases, when used alone each one does not increase efficiency considerably or at all. However, these practices interact in a non-additive way such that employing at least two in combination can dramatically relax the constraints typically associated with field experiments with survey outcomes; in some examples, they decrease variable costs[1] by 98%.

Figure 1 previews some of our results about how these practices can interact in common settings. The Figure considers the variable costs of conducting a study in a common setting in the literature, an experiment studying the persuasive effect of door-to-door canvassing of registered voters in the US that measures outcomes in two rounds of post-treatment surveys, to measure both short-run and long-run effects. Each row in Figure 1 corresponds to a different possible design, all sixteen permutations of using or not using each of the four practices we study. The length of each bar corresponds to the cost of each possible design for achieving a fixed level of precision (a standard error of 0.045 standard deviations), assuming empirical parameters about survey costs and so forth estimated from two empirical studies.[2] The blue bar shows the variable costs of a traditional experiment employing the modal design in the literature, which employs none of the four practices we study and relies on a telephone survey (denoted T) instead of an online survey (denoted O) to

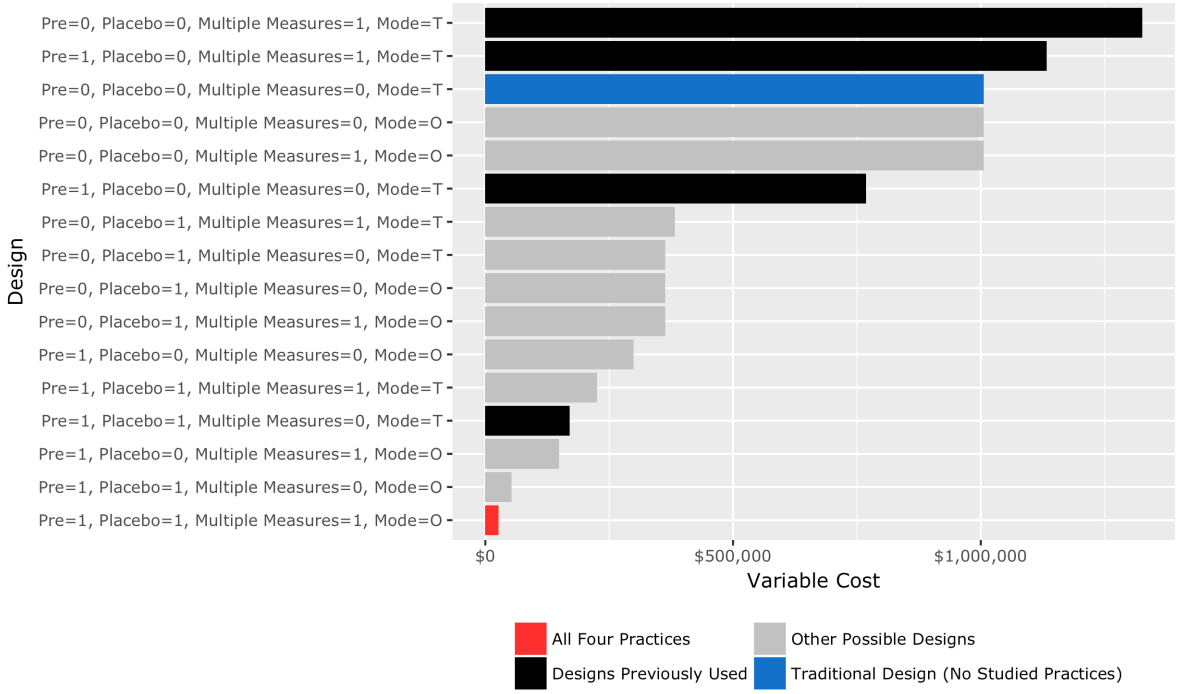
[1]Throughout we consider the variable costs of experiments only, not fixed costs such as the costs of pre-testing a survey instrument, purchasing data on voters, etc.

[2]These parameters are examples only. We describe how we calculated them from our empirical studies and the literature in Online Appendix B.

collect outcomes. The black bars show costs for other designs from existing literature, which have employed some of these practices but rarely multiple in combination. The gray bars show other permutations of designs that might be possible if one were to employ different subsets of these practices. Finally, the red bar shows the variable cost of an experiment using all four practices.

Figure 1 shows that, in this common setting, an experiment using all four practices can be significantly more feasible than an experiment using only one of these practices. An experiment with a variable cost of over $1,000,000 with none of these practices could instead cost approximately $20,000. In addition, such an experiment would be able to precisely test additional design assumptions and require real-world intervention on only a minuscule scale.

Figure 1: Comparing Feasibility of Different Designs



Notes: This Figure uses the framework we developed to estimate the feasability of multiple potential experimental designs for an example door-to-door canvassing study that assumes the empirical parameters described in Online Appendix B.

Of course, Figure 1's empirical results about the benefits of these four practices are specific to

a particular intervention, population, and context. Accordingly, this paper's second contribution is a general and extensible framework that allows researchers to select the most efficient mix of these practices in a wide variety of applications and that can be easily extended to accomodate unique features of particular settings. This framework analytically captures the effect of parameters such as survey response rates, treatment application rates, and the stability of survey responses on the cost of field experiments with survey outcomes that do or do not employ each of the four practices we consider. This framework also captures the gains in efficiency that can arise from the complementarities between the four practices we study. We provide several examples of how researchers can use this framework to select more efficient, robust, and ethical designs in a wide variety of applications, just as Figure 1 did for US door-to-door canvassing study.

Our third contribution is new empirical studies that examine how these practices perform in practice. A first empirical study examines the representativeness of the samples that can be recruited with the survey mode we study: online surveys recruited from a defined sampling frame and surveyed online at least twice. This first study recruited a sample in this manner and compared it to a common approach. In particular, we recruited US registered voters to two rounds of online surveys by mail and compared this sample's representativeness to a sample of the same recruited by the traditional means in existing literature, telephone. Although different recruitment methods may yield different results in different settings, results of this study suggest this recruitment strategy compares favorably to common practice. A second empirical study successfully deploys all four practices in the context of an original field experiment measuring the effects of a door-to-door canvassing effort targeting abortion attitudes. This study estimated a precise (null) effect, indicating it is practical to implement all four of these practices together and that doing so appears able to evade a variety of potential biases.

We conclude by discussing remaining limitations and potential extensions.

2 Field Experiments With Survey Outcomes: Typical Designs And Their Challenges

2.1 Designs Common In The Literature

How do political science researchers typically conduct field experiments with survey outcomes today? Table 1 catalogues the existing, publicly available political science field experiments with survey outcomes of which we are aware.

Table 1 establishes the novelty of this paper's first contribution, which notes the complementarities between the four practices we study. In particular, the Table shows that existing experiments rarely take advantage of the complementarities between these practices. These practices are: 1) surveys administered online to a sample recruited from a list sampling frame of the target population units (Cheung 2005, e.g., a public list of registered voters, a civil registry, membership lists of an activist group, etc.), 2) with at least one baseline wave prior to treatment, 3) with multiple measures of outcomes gathered at each wave analyzed as an index and, if possible, 4) a placebo control. The middle four columns in Table 1 record whether each existing study uses each of these practices. They show that each of these individual practices only occasionally appears in field experiments with survey outcomes. Moreover, these design features rarely appear together in the same field experiment, except in one study we have conducted using this paper's ideas (Broockman and Kalla 2016).

To build familiarity with existing practice, Figure 2 depicts the modal design of the field experiments in Table 1a. Such experiments could employ all of the four practices we study but employ none of them. We will call this 'the traditional design.' An analyst first defines a sample of individuals and randomly assigns them to treatment and control groups. Delivery of the treatment is attempted to treatment group subjects, but many treatment group subjects are not successfully treated. Control group subjects are not contacted. All subjects originally assigned to either condi-

Table 1: Existing political science field experiments with survey outcomes

(a) Placebo possible

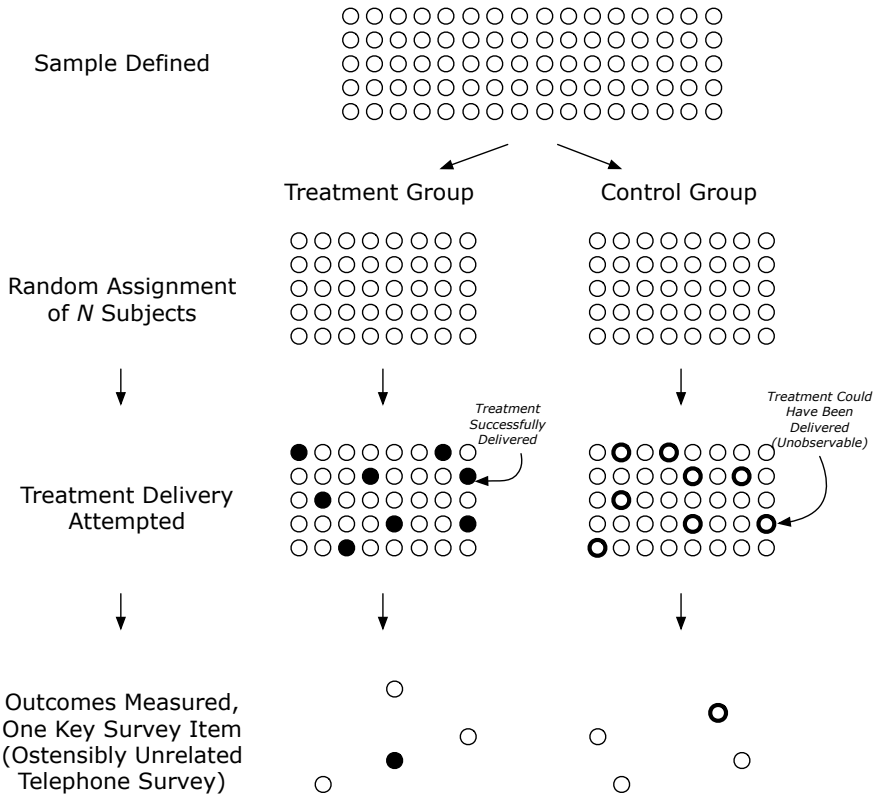
Study	Survey Mode	Baseline Survey?	Index of Multiple Measures?	Placebo?	Multiple Follow-Ups?
Adams and Smith (1980)	Phone	No	No	No	No
Arceneaux (2007)	Phone	No	No	No	No
Arceneaux and Kolodny (2009*a*)	Phone	No	No	No	No
Arceneaux and Kolodny (2009*b*)	Phone	No	No	No	No
Arceneaux and Nickerson (2010)	Phone	No	No	No	No
Barton, Castillo and Petrie (2014)	Phone	No	No	No	No
Bailey, Hopkins and Rogers (2016)	Phone	No	No	No	No
Broockman and Kalla (2016)	Voter File → Online	✓	✓	✓	✓
Cardy (2005)	Phone	No	No	No	No
Dewan, Humphreys and Rubenson (2014)	Phone	✓	No	✓	No
Lam and Peyton (2013)	Phone	✓	No	No	No
Nickerson (2005*a*)	Phone	No	No	No	No
Nickerson (2007)	Phone	✓	No	No	No
Potter and Gray (2008)	Phone	No	No	No	No
All Four Practices	Defined Frame → Online	✓	✓	✓	✓

(b) Placebo not possible

Study	Survey Mode	Baseline Survey?	Index of Multiple Measures?	Placebo?	Multiple Follow-Ups?
Adida et al. (2016)	FTF → Phone	✓	No	n/a	No
Albertson and Lawrence (2009)	Phone	✓	No	n/a	✓
Broockman and Green (2014)	Phone	No	No	n/a	No
Broockman and Butler (2016)	Phone	✓	No	n/a	No
Conroy-Krutz and Moehler (2015)	FTF	No	No	n/a	No
Cubbison (2015)	Phone	No	No	n/a	No
Doherty and Adler (2014)	Phone	No	No	n/a	✓
Enos (2014)	FTF → Online	✓	No	n/a	✓
Fearon, Humphreys and Weinstein (2009)	FTF	✓	No	n/a	No
Gerber (2004)	Phone	No	No	n/a	No
Gerber, Karlan and Bergan (2009)	Phone	No	✓	n/a	No
Gerber, Huber and Washington (2010)	Phone	✓	✓	n/a	No
Gerber et al. (2011)	Phone	No	No	n/a	✓
Humphreys and Weinstein (2012)	FTF	✓	No	n/a	✓
Miller and Robyn (1975)	Phone	✓	No	n/a	✓
Rogers and Nickerson (2013)	Phone	No	No	n/a	No
Sadin (2014)	Phone	No	No	n/a	No
Shineman (2016)	Opt In → Online	✓	✓	n/a	No
Strauss (2009), Section 5.5.4	Phone	No	No	n/a	No
All Three Possible Practices	Defined Frame → Online	✓	✓	n/a	✓

tion are then solicited for an ostensibly unrelated follow-up survey, which few answer, that contains one key survey item of interest.

Figure 2: 'The Traditional Design'



2.2 Challenges Field Experiments With Survey Outcomes Often Face

In this section we review challenges field experiments with traditional designs often face. The following section will formalize how the methodological practices we describe can ameliorate each, especially when used in combination.

To help illustrate key ideas, throughout we assume several example values for marginal costs of surveys, treatment, etc. Online Appendix B describes how we calculated these example values from our empirical studies and the literature. However, we caution readers that these example values are for exposition purposes only and likely vary across contexts and time.

2.2.1 Failure to Treat

Failure to treat arises when some treatment group subjects are not successfully administered treatment. It increases necessary sample sizes (Gerber and Green 2012). To appreciate how, imagine planning an experiment to assess the impact of a door-to-door canvassing treatment powered to detect a 5 percentage point effect. Further suppose canvassers contact 20% of treatment group subjects (as in Bailey, Hopkins and Rogers 2016). A 5 percentage point effect among those contacted would manifest as an overall difference of $5 \times 0.20 = 1$ percentage point between the entire treatment and control groups. A final sample of approximately 80,000 survey responses would be required to detect this 1 percentage point effect with 80% power.

The budgetary implications of failure to treat are especially unfavorable in field experiments with survey outcomes because it increases both the number of subjects one must treat and the number of subjects one must survey. Consider the example just discussed hoping to yield 80,000 survey responses for analysis. Assuming for the moment that survey response rates are 100%, the experimenter must pay to knock on the doors of the 40,000 subjects in the treatment group and to survey all 80,000 subjects. At marginal costs of $3 per canvass attempt and $5 per survey response, the experiment's variable cost would be $520,000, of which $400,000 is survey costs. However, if all subjects in the treatment group could be actually treated, only 3,200 subjects would be necessary, resulting in variables costs of only $20,800, with only $16,000 in survey costs.

2.2.2 Survey Non-Response

Field experiments with survey outcomes usually collect outcomes by telephone, and response rates to telephone surveys in the United States and other developed countries are now typically under 10% (Kohut et al. 2012). In anticipation of this non-response, analysts must treat many more subjects, increasing treatment costs. To see how, consider the experiment described above. Anticipating a response rate of 10% to a final survey, an analyst must attempt to canvass 400,000 voters

in order to yield 40,000 voters *both attempted for canvassing and then successfully surveyed*. Assuming marginal treatment costs scale linearly, this would increase treatment variable costs from $120,000 to $1,200,000 (in addition to the $400,000 in survey costs already discussed).

2.2.3 Limited Pre-Treatment Covariates Available

Finally, many field experiments with survey outcomes have few pre-treatment covariates available that predict outcomes well. For example, Bailey, Hopkins and Rogers (2016) found that commercial scores and administrative data could only predict survey responses to a presidential vote choice question with an R^2 of 0.005. Such limited predictive power has several disadvantages. First, although baseline covariates can increase the precision of estimates (e.g., Sävje, Higgins and Sekhon 2016), covariates that predict outcomes poorly do not meaningfully do so. For example, when $R^2 = 0.005$, the sample size necessary to achieve the same precision decreases by only 0.5%. In addition, lacking prognostic covariates makes differential attrition difficult to detect. Differential attrition arises when the treatment influences survey response rates,[3] leading the surveyed treatment and control groups to differ in expectation even if the treatment has no effect (Gerber and Green 2012, ch. 7). Any experiment with survey outcomes without prognostic pre-treatment covariates cannot persuasively evaluate the assumption of no differential attrition, even though this assumption has been found to fail (Bailey, Hopkins and Rogers 2016). Finally, the absence of pre-treatment covariates also precludes testing many theories with predictions about how treatment effects are moderated by prior attitudes or previous exposure (e.g., Druckman and Leeper 2012).

3 Deriving A Framework for Selecting Experimental Designs

The practices we study are able to substantially ameliorate many of these challenges. In this section we provide a formal analysis comparing the asymptotic efficiency of experiments that employ some

[3] For example, suppose pro-Clinton phone calls discourage Trump supporters from answering surveys later.

or all of the practices we consider to the traditional design shown in Figure 2. We first describe and consider the trade-offs each of these practices involves and how each practice complements the others. We then use these analyses to build a framework for evaluating trade-offs between possible designs using different mixes of these four practices.

Our framework can accommodate a wide variety of possible settings, and all four of the practices we study will not be optimal in all these settings. However, to build understanding about how each of these practices logistically functions, we begin by describing a possible study using all four practices in the setting of a door-to-door canvassing experiment targeting US registered voters' attitudes, just like many existing studies reported in Table 1a and our application study. First, a researcher would send mail to a sampling frame of registered voters inviting them to complete a baseline online survey with multiple measures of outcomes. The survey collects respondents' email addresses so that they can be invited to follow-up surveys later. Next, treatment is delivered to baseline survey respondents only, as is a placebo if possible. Only respondents to the baseline survey are targeted with a real-world intervention ostensibly unrelated to the survey. For example, a canvasser may visit baseline survey respondents' homes and deliver either the treatment or placebo. Finally, the researcher conducts a follow-up survey, but only of individuals who were contacted. Respondents are invited via email to complete these follow-up surveys. Appendix Figure A1 depicts this example design, with the practices we study noted in red.

3.1 Setup for Formal Analysis

In this subsection, we detail the assumptions and estimators that form the basis of our formal analysis of the four practices we study. Readers familiar with the design and analysis of experiments with non-compliance may wish to skip this subsection.

We assume a random sample of size N from an infinite population. Let $z_i \in \{0, 1\}$ denote the treatment randomly assigned to subject i, and let $d_i(z) \in \{0, 1\}$ indicate whether subject i is actually treated when the treatment assignment $z_i = z$. Let $Y_i(z, d)$ denote the potential outcome

for subject i when $z_i = z$ and $d_i = d$. We assume the usual noninterference assumption, so the potential outcome of i only depends on the treatment subject i is assigned. We also make the usual exclusion assumption, $Y_i(z, d) = Y_i(d)$. We define $Y_i(z = 1) = Y_i(z = 1, d = d_i(1))$. Compliers are those subjects who take treatment when they are assigned to the treatment group, and do not take treatment when they are assigned to the control group—i.e., subjects for whom $d_i(1) = 1$ and $d_i(0) = 0$. We assume no subjects assigned to control are treated, such that $d_i(0) = 0$ for all i.

Our estimand of interest is the Complier Average Causal Effect (Gerber and Green 2012, p., 142) defined as:

$$\text{CACE} \quad = \quad \mathbb{E}\left[Y_i(d = 1) - Y_i(d = 0) \mid d_i(0) = 0, d_i(1) = 1\right]. \tag{1}$$

An alternative estimand, which ignores compliance, is the intent-to-treat estimand defined as:

$$\begin{aligned} \text{ITT} &= \mathbb{E}\left[Y_i(z = 1) - Y_i(z = 0)\right] \\ &= \mathbb{E}\left[Y_i(z = 1, d(1)) - Y_i(z = 0, d(0))\right]. \end{aligned}$$

The intent-to-treat effect of treatment assignment (z) on compliance (d) is defined as:

$$\text{ITT}_d \quad = \quad \mathbb{E}\left[d_i(1) - d_i(0)\right],$$

which equals $\mathbb{E}\left[d_i(1)\right]$ because $d_i(0) = 0$ for every i with one-way non-compliance.

With this setup, CACE can be estimated in two ways. First, we can observe who in the control group could have been treated with the placebo design (Nickerson 2005*b*). Observing $d(1)$ for all i, we can plug in sample estimates in Equation 1. We refer to this estimator as $\widehat{\text{CACE}}_{Placebo}$.

The second approach, more common in existing field experiments with survey outcomes and

non-compliance, is:

$$\widehat{\mathrm{CACE}}_{\mathrm{ITT}} \;=\; \frac{\mathrm{ITT}}{\mathrm{ITT}_d},$$

which motivates the usual instrumental variables estimator. As with all field experiments with survey outcomes and non-compliance, both estimates are local to compliers who complete surveys (an issue we return to below).

One may use the delta method to obtain the following asymptotic variance for $\widehat{\mathrm{CACE}}_{ITT}$:

$$\mathbb{V}(\widehat{\mathrm{CACE}}_{\mathrm{ITT}}) \;=\; \frac{1}{\widehat{ITT}_d^2}\,\mathbb{V}(\widehat{ITT}) + \frac{\widehat{ITT}^2}{\widehat{ITT}_d^4}\,\mathbb{V}(\widehat{ITT}_d) - 2\frac{\widehat{ITT}}{\widehat{ITT}_d^3}\,\mathbb{C}(\widehat{ITT}, \widehat{ITT}_d), \qquad (2)$$

where $\mathbb{C}$ denotes covariance.

Prior work in this literature has examined the asymptotic variance of estimators of $\widehat{\mathrm{CACE}}_{\mathrm{ITT}}$ assuming that the estimate of ITT_d is fixed and hence ignoring the last two terms of Equation 2 (Gerber and Green 2012; Nickerson 2005*b*):

$$\mathbb{V}(\widehat{\mathrm{CACE}}_{\mathrm{ITT}}) \;\approx\; \frac{1}{\widehat{ITT}_d^2}\,\mathbb{V}(\widehat{ITT}). \qquad (3)$$

For our purposes, ignoring the last two terms in Equation 2 allows for a cleaner comparison between the variance of $\widehat{\mathrm{CACE}}_{\mathrm{ITT}}$ and the variance of $\widehat{\mathrm{CACE}}_{Placebo}$. As previous authors have noted, these last two terms make little difference in practice. Indeed, the variance of traditional experiments relying on $\widehat{\mathrm{CACE}}_{\mathrm{ITT}}$ is often actually slightly larger than given in Equation 3, making our comparative statements about the efficiency of $\widehat{\mathrm{CACE}}_{Placebo}$ more conservative.[4]

We also make several additional assumptions throughout to simplify exposition of the key

[4]For example, Green, Gerber and Nickerson (2003) report six GOTV experiments. In our analysis of all six, ignoring the last two terms results in slightly smaller variance estimates: the mean ratio of Equation 3 over Equation 2 is 0.994 across them. Other researchers have also observed that the additional terms are very small (e.g., Angrist 1990; Bloom et al. 1997; Heckman, Smith and Taber 1994).

ideas. We assume a balanced experimental design with 50% allocation to treatment and 50% allocation to control. We further assume that there is a constant treatment effect, and hence that the true variance of the potential outcomes in the subject pool is the same for treated and control subjects ($\mathbb{V}[Y(0)] = \mathbb{V}[Y(1)] = \sigma^2$). For simplicity, this variance is assumed to be 1.

Given these simplifications:

$$\mathbb{V}(\widehat{\text{CACE}}_{ITT}) \approx \frac{4\sigma^2}{NA^2}, \tag{4}$$

where N is the number of subjects randomly assigned, and A is the application or contact rate (ITT_d).

3.2 How The Four Practices Can Increase Efficiency

We next formally analyze how each of the four practices we discuss can increase efficiency individually and together. Table 2 verbally summarizes our points. It discusses the primary advantages of each of these practices as it has previously been understood, our results about the special benefits each practice can have in field experiments with survey outcomes, and our results about how each practice can complement others in field experiments with survey outcomes to yield additional improvements.

For our formal analysis of how each of these four practices can increase efficiency, we will consider how an experiment's variable costs $c_{P,B}(\cdot)$ vary with different design choices. $P \in \{0, 1\}$ indicates whether the placebo is used and $B \in \{0, 1\}$ indicates whether a baseline survey is used. Variable cost $c(\cdot)$ is a function of many variables. To reduce notational clutter, we exclude irrelevant variables in each instance and let the context dictate the parametrization. We focus on how variable cost varies as a function of the required sample size N or a desired variance V^*, the number of rounds of post-treatment follow-up surveys one wishes to conduct F (e.g., for $F = 2$ if one wants to test both whether there is an initial treatment effect and then whether any effect lasts in a subsequent round of surveying), and when considered, the marginal cost of attempting treatment

Table 2: Potential benefits of and complementarities between four methodological practices

Methodological Practice	Previously Documented Benefits	Special Benefits in Field Experiments with Survey Outcomes	Benefits Complementing Other Practices (Decreasing Costs or Increasing Benefits of Other Practices)
Placebo (if applicable)	• Identifies compliers in the control group, facilitating estimation of the CACE with much greater precision than the intent-to-treat estimator and meaning fewer individuals must be treated or surveyed to attain the same precision (Nickerson 2005*b*).	• Identifies non-compliers in both groups, allowing non-compliers to be excluded from re-interviews, reducing survey costs.	• Increased precision reduces sample size required for **baseline survey** as well.
Baseline Survey	• Measures covariates at baseline capable of decreasing sampling variability and allowing theories with predictions for heterogenous effects to be tested (Gerber and Green 2012; Bloniarz et al. forthcoming).	• Identifies and establishes a relationship with subjects who can then be reliably re-interviewed, decreasing wasted treatment effort on non-measurable subjects. • Pre-treatment outcomes allow sensitive tests for differential attrition.	• Identifying subjects who can be reliably re-interviewed also reduces the necessary number of placebo interactions, thus decreasing the cost of adopting the **placebo** design. • Allows ones to determine if the compliers are the same in treatment and placebo on observed characteristics; decreasing the risk associated with the **placebo** design.
Multiple Measures Combined Into Index	• Reduces measurement error (Ansolabehere, Rodden and Snyder 2008), increasing the value of every observation and reducing the sample size required.		• Increases the test-retest correlation between the **baseline survey** and follow-up survey, allowing the baseline survey to decrease sampling variability more strongly.
Online Survey Mode	• Allows for additional item formats (e.g., the IAT) and may decrease social desirability bias (Gooch and Vavreck 2016).		• Higher reinterview rates than telephone surveys, strengthening the **baseline survey**'s ability to identify follow-up respondents. • **Multiple measures** can be included less expensively and with less suspicion, decreasing measurement error.

T and conducting a survey S.

Note that these parameters are examples only and could vary dramatically in different settings. See Online Appendix Section B for details on how we calculated these example parameter values. Moreover, note that we consider variable costs only, and do not take into account fixed costs such as purchasing voter lists, training canvassers, renting office space, or travelling to a country to conduct an experiment.

Table 3 will keep track of notation and the parameter values from our empirical studies we will use in our examples.

3.2.1 Practice 1: Placebo

If failure to treat can occur and be observed, a placebo condition can increase efficiency dramatically (Nickerson 2005*b*). In an experiment with a placebo condition, subjects in the control group are contacted with an unrelated appeal. The purpose of these placebo contacts is to identify control subjects to whom treatment could be delivered – that is, to identify whether control group subjects are compliers or non-compliers. For example, in our second empirical study, canvassers contacted individuals in placebo households about recycling.[5] Subjects in each group who open the door and identify themselves before either regime begins are then used at the basis for comparison when estimating the CACE.

The variance of the CACE estimator with the placebo design is:

$$\mathbb{V}(\widehat{\text{CACE}}_{Placebo}) = \frac{4\sigma^2}{NA}, \tag{5}$$

where A is the fraction of the N subjects who are contacted, such that NA is the number of contacted subjects whose outcomes are compared during estimation. As Nickerson (2005*b*) shows,

[5]The particular placebo used may vary depending on the application. For example, Dewan, Humphreys and Rubenson (2014) use a placebo in which canvassers simply provided information on the date of a referendum while the treatments provided persuasive arguments on the referendum.

Table 3: Notation and Values Used in Examples

Notation	Definition	Value Used In Examples
	Design Parameters	
σ^2	True variance of potential outcomes	1
V^*	Target variance of a prospective study	0.002
F	Number of rounds of post-treatment follow-up surveys	2
	Treatment Parameters	
N	Number of subjects assigned to treatment and control or placebo in total, with $\frac{N}{2}$ assigned to each condition	
A	Proportion of subjects attempted for treatment that are successfully treated	$\frac{1}{4}$
T	Marginal cost of attempting treatment or placebo contact	\$3
	Survey Parameters	
$S_{\text{Mode}\in\{O,T\},\text{Measures}\in\{S,M\}}$	Marginal cost of completed survey; with either Online or Telephone mode and Single or Multiple measures	\$5, except $S_{T,M} = \$10$
$R_{\text{Wave}\in\{1,2\},\text{Mode}\in\{O,T\}}$	Response rate to a first (1) or second (2) round of surveys, collected Online (O) or by Telephone (T). A first round of surveys could refer to a baseline survey before treatment or an endline survey after treatment when there has been no baseline survey. A second round implies only subjects who answered a first round of surveys are solicited.	$R_{1,O} = 0.07$, $R_{1,T} = 0.07$, $R_{2,O} = 0.75$, $R_{2,T} = 0.35$
$\rho^2_{\text{Mode}\in\{O,T\},\text{Measures}\in\{S,M\}}$	R^2 of regression of outcome at follow-up on pre-treatment covariates at baseline; with either Online or Telephone mode and Single or Multiple measures	$\rho^2_{O,S} = .25$, $\rho^2_{O,M} = .81$, $\rho^2_{T,S} = .16$, $\rho^2_{T,M} = .33$

$\widehat{\text{CACE}}_{Placebo}$ is unbiased under several assumptions: "(1) the [treatment and placebo] have identical compliance profiles; (2) the placebo does not affect the dependent variable; and (3) the same type of person drops out of the experiment for the two groups."

As previously studied, the benefit of the placebo design is that it can reduce the number of subjects with whom contact must be attempted (Nickerson 2005*b*). To see this advantage, let

T be the marginal cost of attempting to contact a subject to deliver the treatment or placebo (such as the price a paid canvassing firm charges or the opportunity cost of a graduate student's time 'per knock'). Considering only the cost of attempting to treat subjects, the cost of implementing the traditional design in a sample of size N with no placebo and no baseline survey is $c_{P=0,B=0}(N,T) = \frac{1}{2}NT$, as only the $\frac{1}{2}N$ subjects in the treatment group are attempted to be contacted. Suppose an experiment is being planned with the aim of achieving an estimate with variance V^*. Using Equation 4, delivering treatment in the traditional design thus costs $c_{P=0,B=0}(V^*,T) \approx \frac{1}{2} * 4(\frac{\sigma^2}{V^*})(\frac{1}{A^2})T = 2(\frac{\sigma^2}{V^*})(\frac{1}{A^2})T$. In the placebo design, control group subjects are attempted with the placebo contact. Contact is therefore attempted with all N subjects, such that $c_{P=1,B=0}(N,T) = NT$. Using Equation 5, delivering treatment in the placebo design costs $c_{P=1,B=0}(V^*,T) = 4(\frac{\sigma^2}{V^*})(\frac{1}{A})T$. The placebo design is therefore cheaper when $4(\frac{\sigma^2}{V^*})(\frac{1}{A})T < 2(\frac{\sigma^2}{V^*})(\frac{1}{A^2})T$, which reduces to $A < \frac{1}{2}$ (Nickerson 2005*b*).

Less well-appreciated is that a placebo can produce even larger efficiency gains in field experiments with survey outcomes because non-compliers need not be surveyed. Without a placebo, all subjects must be surveyed. Incorporating the cost of surveying, $c_{P=0,B=0}(N,F,T,S) = N(\frac{1}{2}T + FS)$, where F is the number of rounds of post-treatment follow-up surveys and S is the marginal cost of a survey, assuming a 100% survey response rate for now. To achieve an estimate with some desired variance V^*, using Equation 4 reveals that the traditional design will cost $c_{P=0,B=0}(V^*,F,T,S) \approx 4(\frac{\sigma^2}{V^*})(\frac{1}{A^2})(\frac{1}{2}T + FS)$. Supposing an example contact rate of $A = 1/4$, $c_{P=0,B=0}(V^*,F,T,S) \approx (\frac{\sigma^2}{V^*})(32T + 64FS)$. However, with a placebo, "the group receiving the placebo can serve as the baseline for comparison for the treatment group" (Nickerson 2005*b*). This means subjects who are not successfully contacted in the treatment or placebo groups—all non-compliers—do not need to be surveyed. This reduces survey costs. Incorporating the cost of surveying the AN compliers only, $c_{P=1,B=0}(N,F,T,S) = N(T + FAS)$. Using Equation 5, the placebo design will cost $c_{P=1,B=0}(V^*,F,T,S) = 4(\frac{\sigma^2}{V^*})(\frac{1}{A})(T + FAS)$. Again supposing $A = 1/4$, this reduces to $c_{P=1,B=0}(V^*,F,T,S) = (\frac{\sigma^2}{V^*})(16T + 4FS)$. Note that with $A = 1/4$ the placebo reduces the costs

associated with delivering treatment by half ($32T$ to $16T$) but reduces survey costs 16-fold ($64FS$ to $4FS$). With $F = 2$, $T = 3$, and $S = 5$, this is equivalent to an 88% decrease in variable costs.

Illustrating the first way the practices we study can complement each other, a placebo also reduces the costs of baseline surveys by reducing the number of subjects who must be recruited to a pre-treatment baseline if one is used. To see this, suppose a baseline survey of N subjects is conducted before treatment. Let the marginal cost of each baseline survey also be S. The baseline's variable costs thus are NS. The gross variable cost of incorporating a baseline is an increase in costs of $4(\frac{\sigma^2}{V^*})(\frac{1}{A^2})S$ under the traditional design and only $4(\frac{\sigma^2}{V^*})(\frac{1}{A})S$ with a placebo. If $A = 1/4$, a placebo makes the baseline 75% cheaper to implement.

3.2.2 Practice 2: Pre-treatment Baseline Survey

A pre-treatment baseline survey can increase power in two ways. First, and most obviously, baseline surveys can capture pre-treatment covariates that analysts can use to increase precision. This can decrease costs because smaller sample sizes are required to attain a given level of precision. Second, and less obviously, baseline surveys can also decrease *treatment* costs by identifying subjects who are more likely to be interviewed after treatment. If survey response rates are low, many subjects must be treated to yield each survey response for analysis. By identifying and establishing relationships with subjects who can reliably be re-surveyed and only delivering treatment to these subjects, a baseline survey can dramatically reduce wasted effort treating subjects whose outcomes cannot be measured.

To see these advantages we will now incorporate survey nonresponse and pre-treatment covariates into our analysis and consider the differences between a design with or without a baseline survey. For now we will assume a placebo is used and outcomes are collected by telephone survey. First, consider a design using a placebo, a post-treatment telephone survey, and no baseline survey. Let $R_{1,T}$ represent the response rate to the post-treatment telephone survey among the compliers an analyst attempts to survey, where the subscripts indicate subjects are being surveyed for the first

time and by telephone. If N subjects are randomly assigned, then NA compliers are contacted, and then $NAR_{1,T}$ complier-reporters are surveyed via telephone, Equation 5 shows the variance of this design will be:

$$\mathbb{V}(\widehat{CACE}_{P=1,B=0}) = \frac{4\sigma^2}{NAR_{1,T}}. \tag{6}$$

The cost of this design with a placebo, no baseline, and a telephone survey that collects a single outcome measure is:

$$c_{P=1,B=0}(N, F, T, S) = NFAR_{1,T}S_T + NT, \tag{7}$$

where the first term captures the cost of surveying the $NFAR_{1,T}$ subjects who complete the post-treatment telephone survey, which carries a marginal cost S_T for each of F rounds of surveying; NT captures the cost of attempting to contact N subjects with marginal cost of treatment T. Using Equations 6 and 7, to achieve some desired variance V^*, this telephone-based design would cost:

$$c_{P=1,B=0}(V^*, F, T, S) = 4\left(\frac{\sigma^2}{V^*}\right)\left(FS_T + \frac{T}{AR_{1,T}}\right). \tag{8}$$

Note how Equation 8 shows that low response rates to post-treatment telephone surveys $R_{1,T}$ increase the cost associated with treatment.

Now consider the design with a pre-treatment online survey and a follow-up online survey. Let ρ^2 be the R^2 of a regression of the outcome on pre-treatment covariates from the baseline survey and $R_{2,O}$ be the response rate to an online follow-up survey among those who completed a baseline, with subscripts indicating that the follow-up survey is the second time subjects are being surveyed (the first being the baseline) and the online mode (which we will discuss shortly). This design has variance:

$$\mathbb{V}(\widehat{CACE}_{P=1,B=1}) = \frac{4\sigma^2\left(1-\rho^2\right)}{NAR_{2,O}}. \tag{9}$$

The cost of such a study would be:

$$c_{P=1,B=1}(N, F, T, S) = NFAR_{2,O}S_O + NT + NS_O, \tag{10}$$

where S_O is the marginal cost of an online baseline survey and S_O is also the marginal cost of an online follow-up survey.

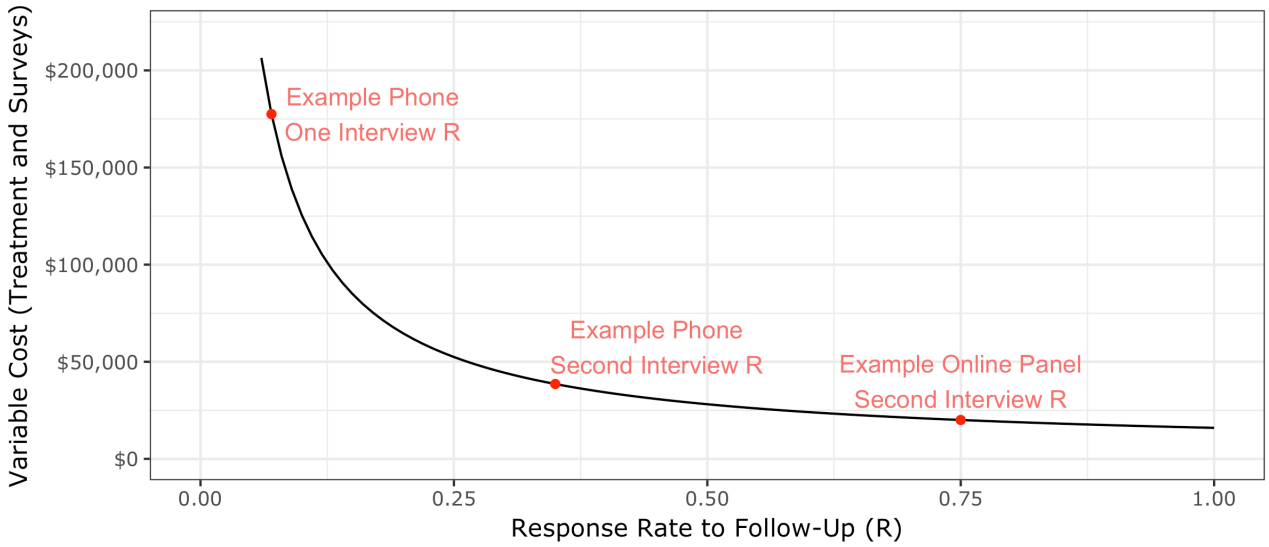
Using Equations 9 and 10, to achieve some desired variance V^*, a design with a baseline survey would cost:

$$c_{P=1,B=1}(V^*, F, T, S) = 4\left(1-\rho^2\right)\left(\frac{\sigma^2}{V^*}\right)\left(FS_O + \frac{T+S_O}{AR_{2,O}}\right). \tag{11}$$

Equation 11 highlights the potential efficiency gains of a baseline survey in two ways. To see these potential gains, compare Equations 8 and 11. First, costs decrease when baseline survey items are prognostic of the ultimate outcome; $(1-\rho^2)$ shrinks the entire cost because the necessary sample size is lower. Second, whereas telephone survey response rates ($R_{1,T}$) are often lower than 10% in developed countries, we have observed response rates to follow-up surveys among those who have already completed baseline surveys ($R_{2,O}$) of about 75% or more (see Online Appendix B). When $R_{1,T} < R_{2,O}$ this reduces the cost of treatment in anticipation that more treated subjects can be surveyed. Figure 3 depicts this latter dynamic. Holding fixed the parameters in Table 3 and varying only the response rate to the follow-up survey, it shows how lower follow-up survey response rates increase costs.

Again illustrating how the practices we study can complement each other, the baseline survey can also dramatically decrease the cost of using a placebo. When a placebo is used but a baseline survey is not, many placebo conversations are wasted on subjects whose outcomes cannot be mea-

Figure 3: How experiment costs decrease with higher survey response rates.



Note: The line is calculated using Equation 11, with parameters except for R from Table 3 held fixed.

sured because they will not complete a phone survey. A baseline survey can reduce placebo costs by reducing the number of placebo conversations wasted on non-responders and, with prognostic pre-treatment covariates, increasing the value of every successful placebo conversation.[6] The ratio of these costs is $\frac{(1-\rho^2)R_{1,T}}{R_{2,O}}$. With the parameter values in Table 3, a placebo costs about 1.8% of what it would cost to implement with traditional designs.

A baseline survey can also help researchers detect or attempt to adjust for differential survey attrition or improper implementation of a placebo.[7]

[6] In particular, with a telephone post-treatment survey only, the cost of placebo conversations was $4(\frac{\sigma^2}{V^*})(\frac{T}{2AR_{1,T}})$. Under the design with a baseline online survey, placebo conversation costs are $4(1-\rho^2)(\frac{\sigma^2}{V^*})(\frac{T}{2AR_{2,O}})$ instead.

[7] As described in Section 2.2.3, differential attrition occurs when the treatment influences who completes a survey. It can bias estimates severely but is often difficult to detect (see Gerber and Green 2012, ch. 7). However, prognostic baseline covariates allow for differential attrition to be detected more sensitively and, if it does occur, for adjustment models to be applied more persuasively (e.g., Bailey, Hopkins and Rogers 2016). Likewise, if a placebo is used, the baseline survey also makes the placebo design less risky to implement because it helps one detect if compliers in each condition differ on baseline outcomes; if implementation of the placebo is found to fail, prognostic baseline covariates may help adjustment models be applied more persuasively.

3.2.3 Practice 3: Multiple Measures Analyzed As An Index

Equation 11 showed how higher test-retest correlations ρ between baseline and outcome measurements increase efficiency. Due to measurement error, one item may have a small correlation between two survey waves even if the underlying attitude it measures is stable. However, when multiple measures of an attitude are collected and combined into an index, stability between survey waves can increase considerably (Ansolabehere, Rodden and Snyder 2008). This increase in stability can increase the precision of estimates dramatically, increasing efficiency.

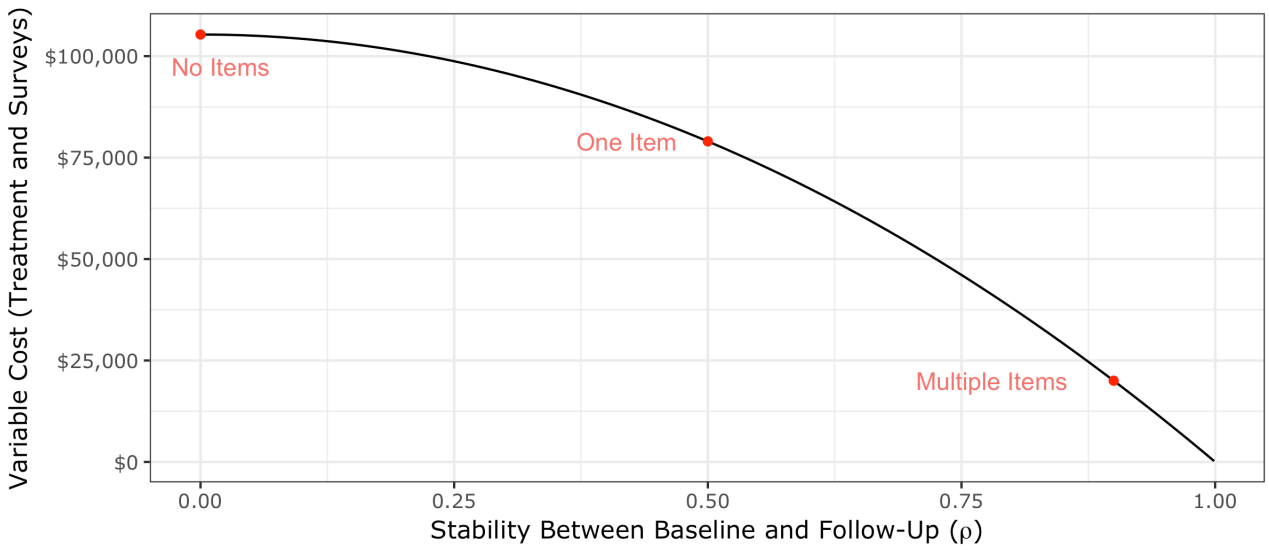
Empirical values from our application study illustrate the magnitude of these potential gains. In that study, analyzing an index of multiple items instead of only one item increases the test-retest correlation ρ to 0.9 from an average of 0.5. This corresponds to a more than three-fold increase to 0.81 from 0.25 for the ρ^2 used in Equation 11, and thus a more than three-fold decrease in costs. Figure 4 shows these gains graphically. Without multiple measures, baseline surveys are less useful for reducing sampling error; the point corresponding to 'One Item' in Figure 3 at $\rho = .5$ implies relatively modest cost savings over a completely unpredictive baseline (at far left). However, with multiple measures, baselines can reduce sampling error tremendously. Note that multiple measures can increase precision even when one item is stable, such as vote choice or partisanship can be; for example, increasing ρ from 0.9 to 0.95 would decrease costs by roughly half.

Although psychology research consistently collects multiple measures to form an index, Table 1 shows that this practice is rare in existing political science field experiments with survey outcomes. We suspect the reason has to do with survey mode, a point to which we turn now.

3.2.4 Practice 4: Online Survey Mode

The fourth practice we study is recruiting individuals to online surveys from a defined sampling frame, such as a list of registered voters (as in our empirical studies), list of all addresses (Jackman and Spahn 2015), FEC donor lists (Barber, Canes-Wrone and Thrower 2016), list of physicians

Figure 4: How experiment costs decrease with higher stability between baseline and follow-up outcome measurements.



The points labeled 'One Item' and 'Multiple Items' are examples only and correspond to the empirical ρs observed in our application study. ρ is calculated by taking the correlation between the follow-up measure and the fitted values from a multivariate regression predicting the follow-up measure using baseline covariates. Note that ρ^2 is the familiar unadjusted R^2 statistic.

(Hersh and Goldenberg 2016), or one of many others (see Cheung 2005). Online surveys can complement the practices studied above in three major ways.

First, online surveys can increase re-interview rates after baseline surveys, increasing R_2. That is, we have observed $R_{2,O} > R_{2,T}$, likely because the first survey can capture additional contact information for each respondent (e.g., an email address) and easily provide them incentives (e.g., a gift card). Increases in R_2 increase the value of baseline surveys. In our work so far, re-interview rates in this mode have sometimes exceeded $R_{2,O} = 80\%$. However, re-interview rates on the phone can be considerably lower; we have observed $R_{2,T} = 35\%$, similar to existing literature (see Online Appendix B).

Second, surveys that collect multiple measures can be cheaper to administer online than by telephone (that is, we have observed $S_{O,M} < S_{T,M}$). For every question in a live telephone survey,

an interviewer must read the question and record respondents' answers. We expect telephone surveys rarely collect multiple measures for this reason. Online surveys rarely carry a high per-question cost. The $5 incentives we have provided for surveys of over 50 questions are much smaller than quotes we received for telephone surveys of this length (see Online Appendix B).

Third, online surveys may have higher test-retest reliabilities, such that $\rho^2_{O,M} > \rho^2_{T,M}$. In our first empirical study we observed larger ρs for the same questions asked online than by telephone.

Such potential increases in R_2 and ρ^2 and decreases in S_M mean collecting outcomes by online panels have the potential to achieve the same precision for less cost than by other survey modes.[8] With this said, two major concerns about online surveys bear mentioning. First, when studies are conducted in other settings, many of these parameter values may change, resulting in different optimal designs (see, e.g., Section 4.3). Second, respondents to online surveys may prove less representative than those recruited with traditional modes. For this reason, we recommend recruiting respondents from an ex ante defined sampling frame. Existing evidence suggests online respondents recruited from a defined frame can be more representative than those who 'opt in' to online surveys (e.g., Brüggen, van den Brakel and Krosnick 2016). More importantly, being able to compare respondents to a defined frame facilitates empirical examination of how representative a sample is on observables. Researchers should also think critically about how unobservable characteristics of those who respond to any survey mode might affect their conclusions. With this said, because the representativeness of subjects recruited to online surveys is a special concern, we will return to this topic with our first empirical study, presented in Section 5.

[8]For example, consider the alternative of phone panels. Using Equation 11, the ratio of treatment and baseline survey costs $N(T + FS)$ for an online panel design and a telephone panel design would be $\frac{(1-\rho^2_{T,S})/R_{2,T}}{(1-\rho^2_{O,M})/R_{2,O}}$. With $\rho^2_{T,S} = 0.16$ for one item in a telephone survey, $\rho^2_{O,M} = 0.81$ for multiple measures in an online survey, $R_{2,T} = 0.35$ for telephone survey reinterview rates and $R_{2,O} = 0.75$ for online survey reinterview rates (see Online Appendix B), this ratio of treatment of survey costs between modes is $\frac{(1-.16)/.35}{(1-.81)/.75} \approx 9$. For the small costs associated with the follow-up surveys, the ratio is $\frac{1-\rho^2_{T,S}}{1-\rho^2_{O,M}} = \frac{1-.25}{1-.81} \approx 4$. Using the parameters from Table 3, the ratio of the total costs is ≈ 8.5. Although exact parameters will vary from study to study, this suggests field experiments that collect outcomes with online survey panels can be nearly an order of magnitude cheaper than field experiments collecting outcomes with other survey modes.

4 A Framework for Selecting Experimental Designs

Scholars wishing to conduct a field experiment with survey outcomes may encounter substantially different design parameters than those explored in the running examples and Table 3. In this section, we provide a framework for how to use the formulas we derived in the previous section to select more efficient and ethical designs. We also provide several examples of how scholars can apply this framework across diverse applications, to their particular questions and setting. These examples will also reinforce our argument that complementarities between these practices can produce large advantages.

Table 4 organizes our analytical results derived in the previous section. As we will show, these formulas allow researchers to compute variances and costs of potential experimental designs as a generic functions of parameters in their settings under alternative permutations of the four design practices we have discussed. The notation in Table 4 corresponds to the same notation defined in Table 3. Subtable 4a gives the variances and costs of alternative designs depending on the presence or absence of a placebo, baseline survey, multiple measures, and online survey mode for cases when compliance can be observed and so a placebo is possible. The presence or absence of placebos and baseline surveys changes these formulas. Survey mode and the presence or absence of multiple measures may change parameters in these formulas but not the formulas. Subtable 4b gives the same but for settings where a placebo is not possible because compliance cannot be observed; these are derived in Online Appendix A.

4.1 Example 1: Door-to-Door Canvassing Study in the United States

Figure 1 at the beginning of the paper previewed how a researcher could use our framework to determine the costs of each of sixteen ways to conduct a door-to-door canvassing study under a given set of empirical parameters. The results in Figure 1 follow from plugging in the parameters from Table 3 to the formulas in Table 4a. In that application, our framework found a design with

Table 4: Variances and variable costs of alternative designs

(a) When placebo possible

Placebo?	Baseline?	$\mathbb{V}(\sigma, \rho, N, A, R)$	$c(N, \cdot)$	$c(V^*, \cdot)$
✓	✓	$\frac{4\sigma^2(1-\rho^2)}{NAR_2}$	$NFAR_2S + NT + NS$	$4(\frac{\sigma^2}{V^*})(1-\rho^2)(FS + \frac{T}{AR_2} + \frac{S}{AR_2})$
✓	No	$\frac{4\sigma^2}{NAR_1}$	$NFAR_1S + NT$	$4(\frac{\sigma^2}{V^*})(FS + \frac{T}{AR_1})$
No	✓	$\frac{4\sigma^2(1-\rho^2)}{NA^2R_2}$	$NFR_2S + \frac{1}{2}NT + NS$	$4(\frac{\sigma^2}{V^*})(\frac{1}{A^2})(1-\rho^2)(FS + \frac{T}{2R_2} + \frac{S}{R_2})$
No	No	$\frac{4\sigma^2}{NA^2R_1}$	$NFR_1S + \frac{1}{2}NT$	$4(\frac{\sigma^2}{V^*})(\frac{1}{A^2})(FS + \frac{T}{2R_1})$

(b) When placebo not possible

Baseline?	$\mathbb{V}(\sigma, \rho, N, A, R)$	$c(N, \cdot)$	$c(V^*, \cdot)$
✓	$\frac{4\sigma^2(1-\rho^2)}{NR_2}$	$NFR_2S + NT + NS$	$4(\frac{\sigma^2}{V^*})(1-\rho^2)(FS + \frac{T+S}{R_2})$
No	$\frac{4\sigma^2}{NR_1}$	$NFR_1S + NT$	$4(\frac{\sigma^2}{V^*})(FS + \frac{T}{R_1})$

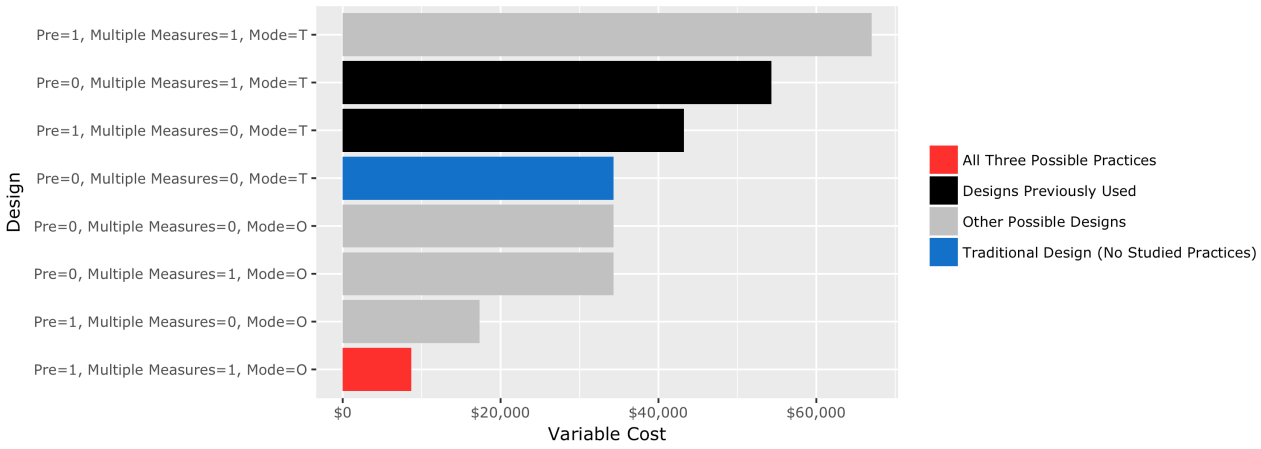
variable costs approximately 98% lower than common designs. In the remainder of this section we show how our framework can be applied to a variety of other settings.

4.2 Example 2: Mailing Information About Members of Congress

In some settings, a placebo is not possible because compliance cannot be observed. Suppose a researcher wants to examine how individuals learn and retain information about their Members of Congress. A researcher might want to include individuals in many Congressional districts to expand the generalizability of the conclusions. A door-to-door canvass treatment would be difficult to deploy on this nationwide basis, but a mail experiment would be practical. However, one cannot easily observe whether a person opens a piece of physical mail, so a placebo could not be used. Subtable 4b gives formulas for alternative designs in situations where a placebo is not possible. To select the optimal design, we will use these formulas and again use the values in Table 3, but substitute $T = \$1$, corresponding to an example mail treatment with a marginal cost of \$1.

Figure 5 provides the results of applying our framework to this experimental design problem. Under these conditions, employing all three possible practices reduces variable costs from approx-

Figure 5: Applying The Framework When Placebo Not Possible: Mail Example



imately $34,285 if none of these practices are used to approximately $8,666. Interestingly, in this application our framework also surfaces that using each of two of these practices alone may actually increase variable costs.

4.3 Example 3: The World Bank Studying A Public Health Intervention in Liberia

Our motivating examples so far have considered how to study the effect of field treatments on political attitudes in the United States, but our framework is much more general. Moreover, it can show how different designs may be more optimal for researchers pursuing different aims in different contexts.

As an example of how our framework can be extended to a different setting, we consider a recent study by The World Bank examining how Ebola infections affected self-reported outcomes such as employment and schooling in Liberia (Himelein 2015). These outcomes were collected in a telephone panel survey.

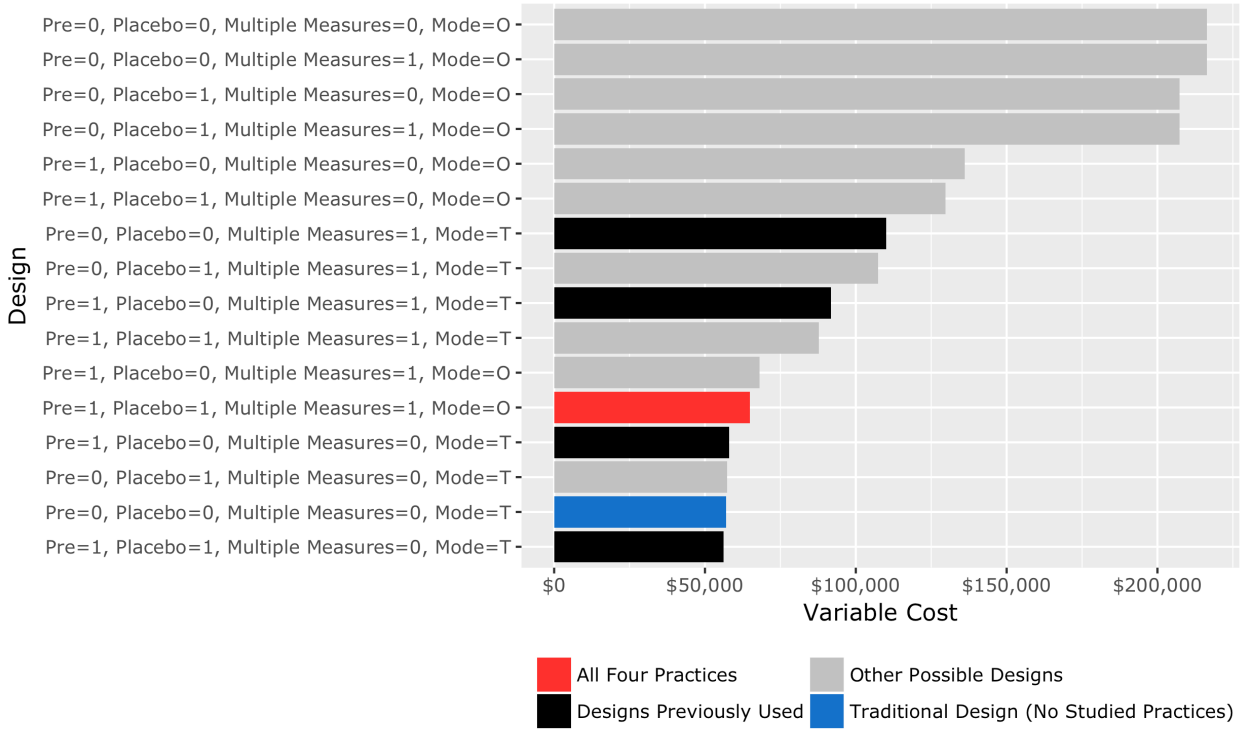
Suppose these researchers wanted to conduct a field experiment in Liberia to estimate the effect

of a public health worker visiting households providing public health information about avoiding Ebola on these outcomes. Our framework is first able to identify the key parameters of interest researchers must forecast to determine which designs would be optimal. In their Liberia study, researchers from the World Bank conducted five rounds of mobile telephone surveys ($F = 5$). The initial survey response rate ($R_{1,T}$) was 28% and the follow-up telephone survey response rate ($R_{2,T}$) was 73%. Indicative of the share of people who can be reached at home in Liberia when one knocks on their door, the contact rate in the face-to-face Afrobarometer survey conducted in May 2015 in Liberia (Isbell 2016) was 97%, so we assume a treatment application rate $A = 0.97$. However, suppose in Liberia an attempted visit from a public health worker is inexpensive given lower wages, such that $T = \$1$, but that online surveys would be much more expensive because many people do not have internet access and would need to be provided it ($S_O = \$20$). For the sake of simplicity, we let online and telephone surveys have the same response rates ($R_{1,T} = R_{1,O}, R_{2,T} = R_{2,O}$) and let V^*, σ^2, ρ^2, and S_T remain unchanged from Table 3.

Figure 6 applies our framework to this setting, examining the most feasible way to conduct this study. In this example, using all four practices we study would not be the most efficient option, nor would the traditional design in the literature. Instead, it would be a telephone survey with a baseline survey and placebo but without multiple measures, to keep the survey short.

The results in Figure 6 could also help these researchers navigate more complicated trade-offs. Suppose a collaborating Non-Governmental Organization refused to implement a placebo condition. The researchers could now detect that conducting a baseline survey is not optimal given that there will be no placebo, even though a baseline survey was optimal when the placebo was present. Alternatively, suppose the researchers wanted to collect multiple measures of outcomes to match existing questionnaires. Our framework now suggests that conducting an online survey may be worth the additional cost, as the parameters we input assumed that a phone survey collecting multiple measures offered less cost savings on marginal costs than a short phone survey. These examples illustrate how our framework can surface subtle complementarities and trade-offs between

Figure 6: Example Results: Variable Costs for Studying Public Health Intervention in Liberia



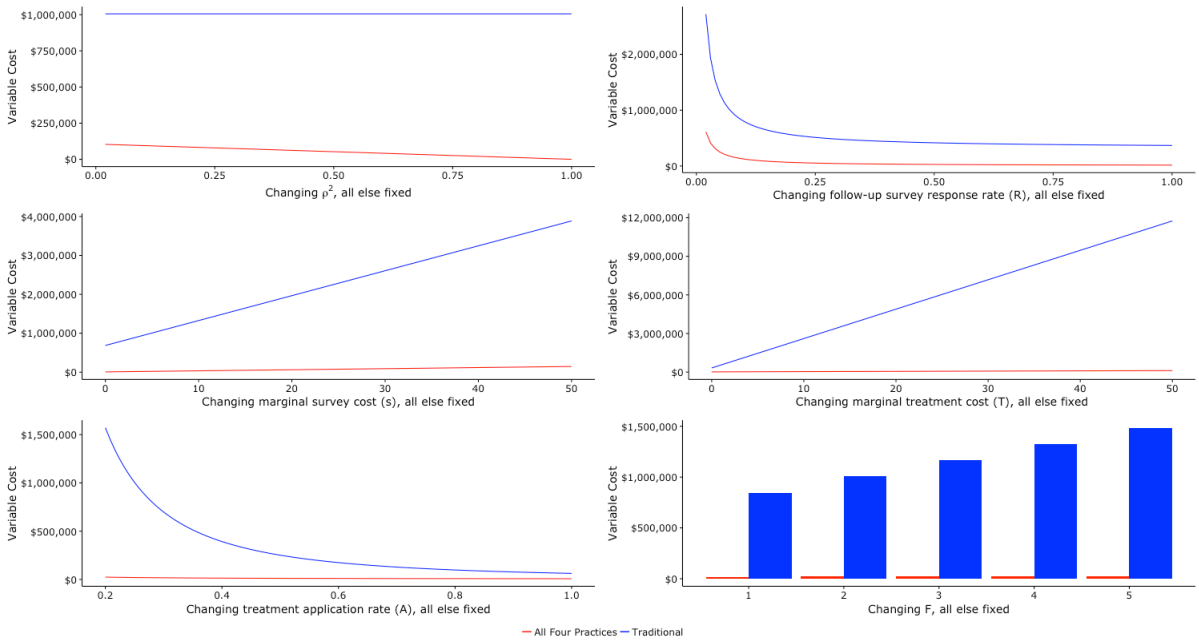
these practices. Our framework also allows researchers to consider more traditional trade-offs. For example, suppose researchers considered using an online survey without providing internet access to those who did not have it, limiting the sampling frame to pre-existing internet users but eliminating the cost of providing internet access. Our framework would allow researchers to compute the money this choice would save and allow them to consider whether this cost savings was worth the potential bias and external validity limitations this would introduce.

4.4 Example 4: Comparing Designs as Parameters Vary

This framework also allows researchers to consider how various design decisions might generalize as design parameters change. For example, what if a researcher has priors over a range of values a parameter may take and wants to test how design decisions may change across that range?

Figure 7 shows how a researcher conducting a door-to-door canvassing study like the one discussed in the running example could use our framework to compute how variable costs would change as each of the parameters the framework considers changes, holding all other parameters at their Table 3 level. In each of the six panels of Figure 7 we vary one of six parameters and show the cost of a study using the literature's traditional design and a design using all four practices. Figure 7 thus demonstrates how our framework can help researchers select superior experimental designs under a wide variety of circumstances. That the design using all the practices we study is consistently superior for these ranges of parameter values also suggests that scholars in a variety of settings may benefit from considering these practices.

Figure 7: How Variable Costs Change with Parameters



4.5 Example 5: Internalizing Ethical Externalities

Many have expressed concern that large field experiments might change collective political outcomes (Michelson 2016). When field experiments require trying to change tens of thousands of in-

dividuals' minds, this concern is especially salient. Our framework is also able to help researchers internalize such potential ethical externalities. Suppose a researcher plans to study a treatment administered by phone that attempts to persuade registered voters on an issue with a marginal cost of attempting treatment T of \$3 (e.g., a 'cost per dial'). Further suppose this researcher perceives the ethical externality of attempting each conversation as approximately \$10. This would increase the marginal cost of attempting treatment T from \$3 to \$13. Re-computing the variable costs of these experiments, the experiment not using any of the four practices would go from approximately \$1,006,000 to \$3,291,400 in variable cost, reflecting an ethical externality of \$2,285,400. The experiment using all four of these practices would go from approximately \$20,015 to \$30,145 in cost, reflecting an ethical externality of \$10,130.

This example establishes two points. First, this example shows the potential advantages in variable cost we have studied can also confer an ethical advantage: because under many conditions using some of these practices means many fewer individuals need to be treated, researchers can reduce the scope of their potential to influence real-world outcomes. More generally, this example shows how our framework can be used to consider a wide variety of potential issues that arise when considering alternative field experimental designs. For example, researchers who assigned a subjective value of \$50,000 to the robustness to differential attrition a baseline survey provides could integrate this value into our framework as well.

5 Empirical Study: Representativeness of US Registered Voters Recruited By Mail To An Online Panel

We now present two empirical studies that examine how the practices we study perform in real applications.[9] First, we examine the representativeness of a sample recruited to an online panel

[9]Replication data for all empirical studies are available at Broockman, Kalla and Sekhon (2017), see doi:10.7910/DVN/EEP5MT.

survey from a defined sampling frame. Scholars may wonder how subjects recruited to one or more rounds of online surveys may differ from subjects recruited by the literature's traditional means, a single round of phone surveys. Our first empirical study considers this issue in detail. Specifically, we recruited US registered voters by mail to two rounds of online surveys and compared their representativeness to that of one and two rounds of phone survey respondents, using both the original sampling frame and other surveys as benchmarks.

We expected online surveys recruited from an ex ante defined frame to yield fairly representative but slightly more educated samples. Debates continue about the generalizability of 'opt in' online survey samples recruited by online ads (e.g., Hill et al. 2007), but research has generally found that surveys administered online are fairly representative when their samples are recruited from ex ante well-defined sampling frames (Brüggen, van den Brakel and Krosnick 2016), with the exception that online samples tend to be slightly more educated on average given that more educated people are more likely to have internet access (Hall and Sinclair 2011). However, existing research that considers the representativeness of online samples typically focuses on particular areas that may yield idiosyncratic results (e.g., Barber et al. 2014; Collins and Rosmarin 2016) and often does not compare results to alternative recruitment methods in the same samples. We therefore sought to gather additional data on this question.

To consider the representativeness of online survey samples empirically, we randomly assigned a random sample of US registered voters to a telephone survey or to an online survey recruited by mail. This allows us to assess the general representativeness of this design on an absolute basis and in comparison to current practice. In addition, we used this nationwide exercise to inform the example parameter values used in Table 3 and our running examples; see Online Appendix B for discussion. (We fully expect these parameters would differ in non-random samples selected for particular studies.)

In early 2016, we purchased a national random sample of the publicly available list of registered voters, observed demographics available on the voter file, modeled demographics available

from our data vendor, their mailing address, and, if available, their landline and mobile telephone numbers. This starting sample provides our first benchmark for representativeness. We then randomly assigned these voters to mail-to-online or phone modes and conducted the surveys. See Online Appendix Section C.1 for details on the data, random assignment procedures, and survey recruitment procedures. One note described further in the Online Appendix to which we want to draw attention is that subjects' race is only observed in some states; it is the product of a statistical model in other states, and therefore we call this variable 'Modeled Race.'

We first compare the administrative data available for the entire sampling frame to the data for just those who completed the baseline ('t0') and follow-up ('t1') online and phone surveys. Figure 8 shows the proportion of various characteristics present in these subsamples. (Table OA2 reports point estimates.) Unsurprisingly, neither online nor phone respondents match the sampling frame exactly on every covariate, nor is either mode superior on every covariate. However, one way to assess the overall representativeness of each sample on observable characteristics is to compute the loss in efficiency that results when each sample is weighted back to the sampling frame, or the design effect (Kish 1965). To calculate the design effects for each mode, we calculated survey weights using entropy balancing (Hainmueller 2012) and logistic regression using gender, modeled race, party identification, 2008, 2010, 2012, and 2014 voter turnout, age, and age squared. Table 5 shows the design effects for each method. A design effect of 1 would indicate perfect representativeness on observables; larger design effects indicate larger differences between the sampling frame and sample on observables.[10] Overall, both online survey waves had smaller design effects than the sample of individuals who responded to either one or two phone surveys.[11] With this said, we only have access to a limited number of variables on the US voter file, and further research with

[10]Individuals without phone numbers were all assigned to the online sample group as they could not be recruited by phone, meaning the online sample eligible universe overrepresents individuals without phone numbers. We adjust the estimated design effects to take into account this overrepresentation. We also only recruited a subsample of the first wave of online survey respondents to the second online survey wave. The estimated design effect for the second online wave also takes this subsampling into account.

[11]In Online Appendix D we present data on the representativeness of subjects in our second empirical study who are recruited by this mode *and* are compliers.

access to other variables would be of interest.

Figure 8: Average administrative data values for sampling frame and respondents.

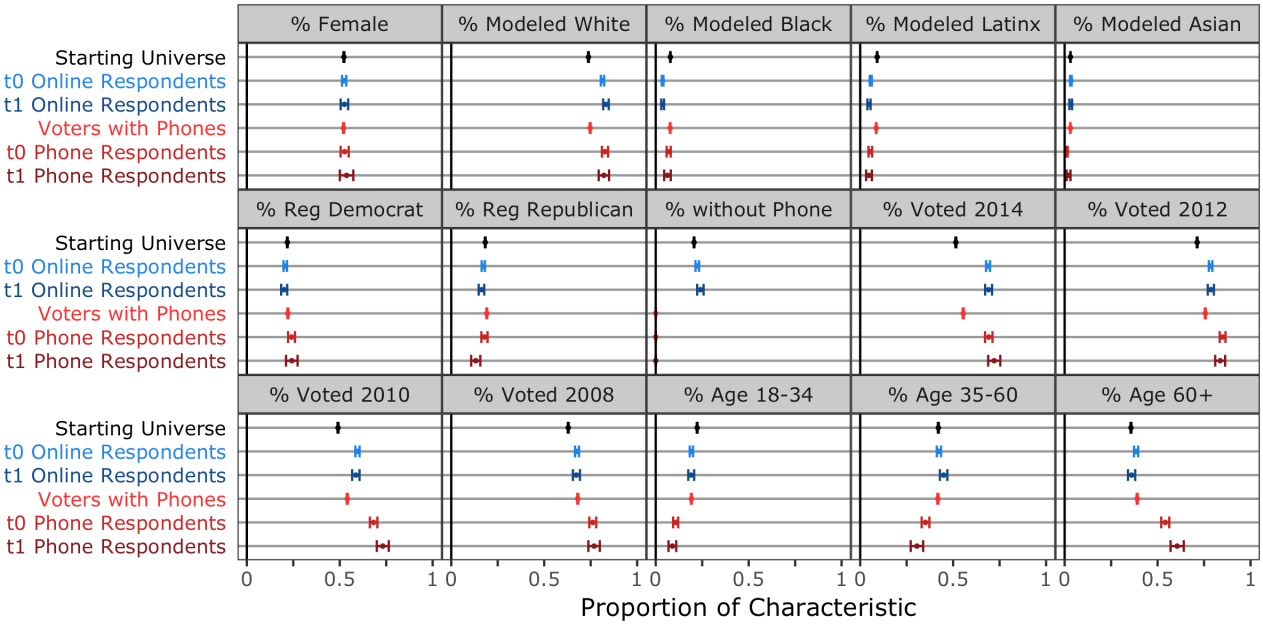


Table 5: Design Effects of Online and Phone Surveys and Panels

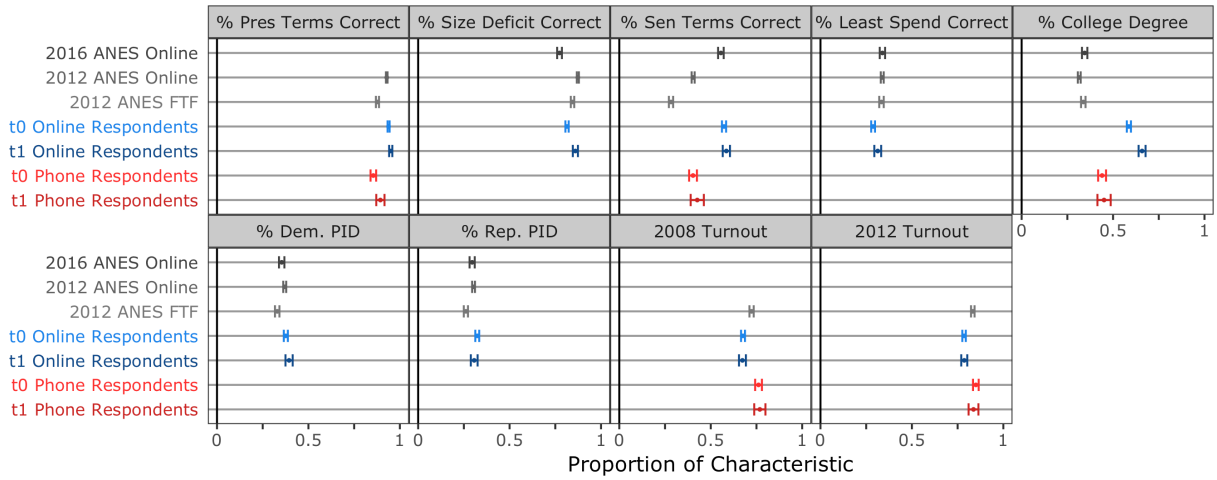
Survey	*Design Effect, Weights From Logistic Regression*	*Design Effect, Weights From Entropy Balancing*
t0 Online	1.23	1.09
t1 Online	1.17	1.17
t0 Phone	1.53	1.38
t1 Phone	1.85	1.71

We also compare these samples to the 2012 ANES and 2016 ANES Pilot Studies in Figure 9, some of which were conducted online as well.[12] We limit the ANES to only registered voters to match the sampling frame from our online survey. We then compare the sample on several

[12]The 2012 ANES used two different recruitment modes: an online survey conducted by GfK Knowledge Networks (denoted '2012 ANES Online' in Figure 9) and a traditional face-to-face survey (denoted '2012 ANES FTF' in Figure 9). The 2016 ANES Pilot Study (denoted '2016 ANES Online' in Figure 9) was conducted online by YouGov in January 2016. We mean to imply no claims on the representativenss of these ANES studies; instead, they serve as a useful comparison familiar to many political scientists.

questions that overlap: political knowledge (size of federal deficit and relative size of U.S. federal spending), party identification, reported education, and validated voter turnout (for the 2012 ANES face-to-face sample).[13] As expected, our online survey respondents are slightly more likely to be well-educated, politically active, and informed, but otherwise generally match the 2012 and 2016 ANES samples. (Table OA3 reports point estimates.)

Figure 9: Covariates collected in 2012 and 2016 ANES, telephone, and online surveys.



Note: Where missing, the question was not asked. Partisanship questions exclude leaners. For the online survey, size of the deficit, spending, education, and partisanship were asked in the first survey wave and the presidential and Senate terms were asked in the second wave. For the phone survey, education and the presidential and Senate term questions were asked in the second wave.

Our finding that online survey respondents in the US appear to be somewhat more educated, active, and informed[14] underscores a broader need for caution for all researchers using field experiments with survey outcomes: researchers should think critically about how respondents to any survey might differ on both observable and unobservable characteristics in their particular setting.

[13] Debates continue about the accuracy of validated turnout in the ANES due to vote file matching issues, so we encourage some caution when interpreting the turnout results; the ANES estimate may be downwardly biased (Berent, Krosnick and Lupia 2016).

[14] Results on several non-political items we asked also reinforce our finding that online survey respondents are slightly more likely to be educated and informed. Specifically, in Online Appendix C.3 we compare the representativeness of the samples to a 2014 Pew survey on scientific knowledge (Funk and Goo 2015) and find that online respondents are slightly more likely to correctly identify answers to scientific knowledge questions.

For example, some theories would predict that relying on estimates from a sample with higher levels of education and information might lead to underestimates of population average treatment effects when studying political persuasion (e.g., Zaller 1992), a fact political scientists studying persuasion with field experiments with survey outcomes should bear in mind. More generally, the representativeness of different survey modes is likely to vary across settings in ways that will be specific to these settings and of which researchers should remain cognizant. By beginning with a defined sampling frame, however, researchers can better empirically examine representativeness on at least observable characteristics. In addition, researchers should be cognizant of whether the mode of their treatment interacts with the mode of their survey; for example, subjects recruited to a phone survey may be especially susceptible to persuasion by phone.

In summary, these data provide cautious optimism that online panel surveys can be capable of recruiting subjects that compare favorably to the representativeness of subjects recruited by phone. However, our findings also reinforce that researchers should think critically about how survey respondents might differ in their particular settings in ways relevant to their research questions.

6 Application Study: Door-to-Door Canvassing on Abortion

In this section we report an original study of a door-to-door canvassing experiment deploying all four of the practices we study. This application study illustrates two main points. First, readers may wonder whether it is logistically feasible to combine some of the practices we study. This application study establishes that an experiment employing all four of these practices is practical to execute and that it does indeed yield the efficiency advantages our framework indicates. Second, this study helps assuage potential concerns that experiments that conduct pre-treatment baseline surveys are especially prone to demand effects. An experiment with a baseline survey involves multiple interactions between researchers and subjects, introducing the possibility that subjects in the treatment group will draw a connection between the surveys and treatment and report on the

surveys what they believe those responsible for the treatment want to hear. However, this study estimated a precise null effect.

During 2015, volunteers from the Los Angeles LGBT Center's Leadership LAB went door-to-door in Los Angeles County seeking to increase support for safe and legal abortion and attempting to reduce stigma towards women who have had abortions. The conversations lasted roughly 10 minutes on average and involved canvassers asking subjects to tell stories about when subjects had made mistakes in their relationships.

We worked with the Los Angeles LGBT Center to deploy an experiment to measure the effects of these conversations that used all four of the practices we have studied. First, the Los Angeles LGBT Center selected LA County neighborhoods that had voted against expanded abortion access in prior ballot initiatives and provided us the publicly available data on registered voters in these neighborhoods. We recruited these voters to an online survey panel via mail sent to the address at which they were registered to vote. 1,982 subjects completed the baseline survey. Most survey items were unrelated to abortion. Next, we randomly assigned respondents to receive an abortion-focused canvass (treatment) or to a recycling conversation (placebo), blocking on an index of baseline responses. Volunteers then knocked on subjects' doors. Regardless of condition, they first identified subjects and marked them as compliers. Canvassers then delivered the treatment corresponding to the subject's random assignment, either the abortion or placebo conversation.[15] Online Appendix D reports intervention details. One week after canvassing occurred, we invited subjects who were successfully reached at the door (compliers) to the follow-up survey via email. We again invited the same subjects to participate in a second follow-up survey five weeks after canvassing took place.

Observed design parameters were consistent with expectations[16] and reinforce the opportu-

[15]Importantly, the survey and canvassers bore different affiliations. The survey was affiliated with UC Berkeley but the volunteers represented the Los Angeles LGBT Center.

[16]The R^2 from regressing abortion attitudes from the first and second post-treatment surveys on pre-specified baseline attitudes and covariates were both 0.81, even though single items had R^2 statistics in the 0.41 - 0.70 range. Response rates to the follow-up surveys were 81% and 79%, respectively.

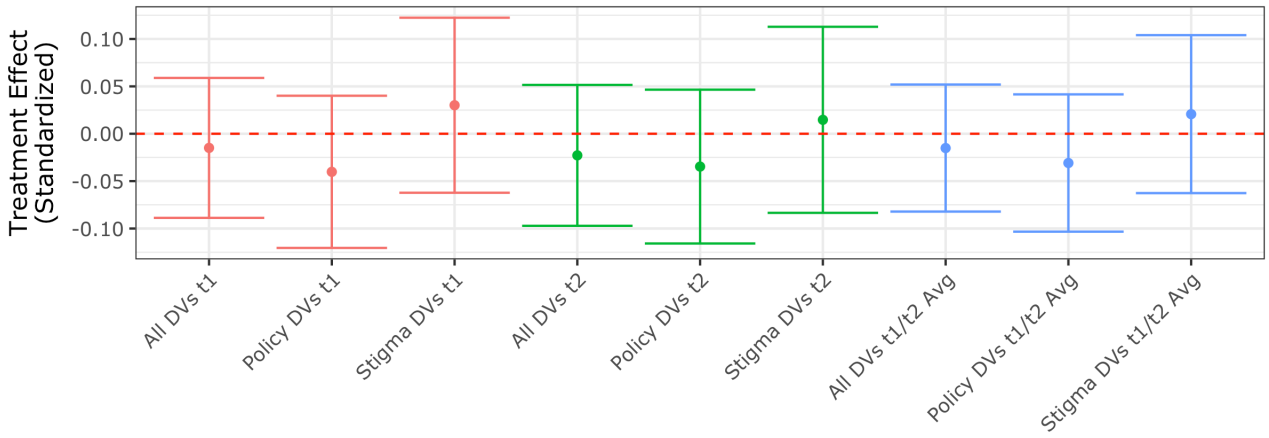
nities for experimentation the four practices we study make available. In all, the surveys cost approximately $16,200 using the design with all four practices, but would have cost approximately $265,000 with traditional designs. The design also only required Los Angeles LGBT Center volunteers to contact approximately 450 voters, but had the same precision as would a study using the traditional design if they had contacted approximately 46,000 voters. To reach 46,000 voters, Center volunteers would have needed to continue canvassing for over a decade at their same pace; our study took them a few days.

To estimate treatment effects on answers to the follow-up survey, we use a linear regression with an index of follow-up survey items as outcomes and the pre-treatment answers to those items as controls. Standard errors are cluster-robust, with clustering at the household level. The outcome indices were computed by taking the first factor from factor analysis and rescaling them to a standard deviation of 1. We show the results for three dependent variables: an index of all the abortion items, an index of just the policy-relevant items, and an index of just the stigma-relevant items. The outcomes are measured for just the first post-treatment survey, just the second post-treatment survey, and an average of both post-treatment surveys (to further reduce measurement error). Online Appendix D.3 gives more detail. These analysis procedures were pre-registered.

Results indicate the treatment effect as a precisely estimated zero on all outcomes: the confidence intervals rule out positive effects of approximately 0.05 standard deviations, which is half the size conventionally considered 'small.' Figure 10 shows these results. Online Appendix D reports balance checks and representativeness assessments.

In addition to demonstrating that these four practices are feasible to deploy in tandem, this study's null result is encouraging for the validity of studies using these practices that find non-null results. One might worry that surveying subjects multiple times about the topic of an experiment necessarily introduces demand effects; subjects may make the connection between the online surveys and the treatment and adjust their survey responses to satisfy the researchers even if their attitudes did not change. This application study's null result suggests these practices are capable

Figure 10: Treatment effect estimates of canvassing on abortion attitudes.



Note: Outcomes are indices of items rescaled to unit variance. Point estimates can be interpreted as standard deviations. 95% confidence intervals surround point estimates.

of precisely estimating null treatment effects.

7 Concluding Discussion

The use of randomized experiments and survey-based research in the social sciences has mushroomed. Together with rising interest in these methodologies, many scholars have begun to conduct field experiments with survey outcomes: experiments where outcomes are measured by surveys but randomized treatments are delivered by a separate mechanism in the real world. However, challenges familiar to experimental researchers and survey researchers – survey non-response, survey measurement error, and treatment non-compliance – mean that common designs for field experiments with survey outcomes are extremely expensive and pose ethical challenges. In this paper, we showed that four practices uncommon in such experiments can yield particularly large gains in efficiency and robustness when they are used in combination. In some settings, the magnitude of these efficiency gains is extremely large. For example, the modal political science field experiment

with survey outcomes in published work is a door-to-door canvassing experiment among registered voters in the United States. But to conduct a well-powered experiment using designs common in the literature, researchers in many settings may well require budgets larger than that of the entire 2016 ANES. However, using all of the practices we study can decrease an experiment's variable costs to a level doctoral dissertation improvement grants could cover.

This paper also developed a framework that will help researchers select the design that is most optimal in diverse settings where treatment costs, survey costs, survey response rates, and other parameters may change. This framework identifies the key parameters that determine an experiment's variable costs and allows researchers to examine the feasibility of a range of possible designs given these parameters. As we discussed, this framework is widely applicable and easily extensible. For example, researchers could use it to internalize the ethical externalities of treating many subjects or quantify the costs of introducing design practices expected to increase robustness. To accompany this paper, we are also making code available that implements this framework.

Although we are optimistic about the potential applications of the practices we study, several open questions remain. First, all experiments with survey outcomes only estimate effects for individuals who both receive the treatment (compliers) and agree to be surveyed (reporters). Data from our application study shown in Figure OA2 suggests complier-reporters are not highly unrepresentative on observables. Nevertheless, generalizing from these local average treatment effects to population treatment effects requires additional assumptions (Hartman et al. 2015). Although these limitations are theoretically similar regardless of the design used, the empirical representativeness of compliers under this design is a clear question for future research. With this said, future research seeking to benchmark the use of survey outcomes in field experiments against behavioral benchmarks (such as precinct-randomized experiments) may be able to take advantage of our framework to reach more precise survey-based estimates for validation.

Second, baseline surveys may have unintended effects that produce bias or reduce external validity. For example, answering survey questions about a topic might change how people later

process information about it, such as by increasing attentiveness (e.g., Bidwell, Casey and Glennerster 2015). Most evidence on this phenomenon is either from developing countries or several decades ago, so it is unclear to what extent present-day populations in developed countries would exhibit such effects. Individuals who answer a survey twice may also be systematically different than those who answer once, as our first empirical study found for phone surveys. This is an important area for future research, with designs readily available in classic psychometric literature (e.g., Solomon 1949).

Answering multiple follow-up surveys after a treatment may also produce biased estimates of treatment effects' persistence over time if subjects remember how they answered particular questions in a previous survey wave. Existing survey and field experiments that track long-term effects (e.g., Coppock 2016) often observe rapid decay in treatment effects, so this bias clearly does not always exist. Refreshment samples or randomly staggered interview times could help address this possibility.

The particular implementation of each of the practices we studied may also be open to improvement. For example, one possible extension to conducting a baseline survey is to conduct multiple baseline waves prior to treatment. Multiple baselines would further increase stability (increasing ρ^2) (McKenzie 2012) and could help identify subjects even more likely to participate again (increasing R). Our framework could be readily applied to determine whether the costs of an additional baseline wave prior to treatment would outweigh these benefits.

We look forward to future research extending these practices and our framework and employing them to shed light on a variety of substantive questions.

References

Adams, William C. and Dennis J. Smith. 1980. "Effects of Telephone Canvassing on Turnout and Preferences: A Field Experiment." *Public Opinion Quarterly* 44(3):389–395.

Adida, Claire, Jessica Gottlieb, Eric Kramon and Gwyneth McClendon. 2016. "How Coethnicity Moderates the Effect of Information On Voting Behavior: Experimental Evidence from Benin." Working Paper.

Albertson, Bethany and Adria Lawrence. 2009. "After the Credits Roll: The Long-Term Effects of Educational Television on Public Knowledge and Attitudes." *American Politics Research* 37(2):275–300.

Angrist, Joshua D. 1990. "ERRATA: Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." *The American Economic Review* 80(5):1284–1286.

Ansolabehere, Stephen, Jonathan Rodden and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102:215–232.

Arceneaux, Kevin. 2007. "I'm Asking for Your Support: The Effects of Personally Delivered Campaign Messages on Voting Decisions and Opinion Formation." *Quarterly Journal of Political Science* 2(1):43–65.

Arceneaux, Kevin and David W. Nickerson. 2010. "Comparing Negative and Positive Campaign Messages: Evidence From Two Field Experiments." *American Politics Research* 38(1):54–83.

Arceneaux, Kevin and Robin Kolodny. 2009*a*. "Educating the Least Informed: Group Endorsements in a Grassroots Campaign." *American Journal of Political Science* 53(4):755–70.

Arceneaux, Kevin and Robin Kolodny. 2009*b*. "The Effect of Grassroots Campaigning on Issue Preferences and Issue Salience." *Journal of Elections, Public Opinion and Parties* 19(3):235–249.

Bailey, Michael A., Daniel J. Hopkins and Todd Rogers. 2016. "Unresponsive, Unpersuaded: The Unintended Consequences of Voter Persuasion Efforts." *Political Behavior* .

Barber, Michael J., Brandice Canes-Wrone and Sharece Thrower. 2016. "Ideologically Sophisticated Donors: Which Candidates Do Individual Contributors Finance?" *American Journal of Political Science* .

Barber, Michael J., Christopher B. Mann, J. Quin Monson and Kelly D. Patterson. 2014. "Online Polls and Registration-Based Sampling: A New Method for Pre-Election Polling." *Political Analysis* 22(3):321–335.

Barton, Jared, Marco Castillo and Ragan Petrie. 2014. "What Persuades Voters? A Field Experiment on Political Campaigning." *The Economic Journal* 124(574):F293–F326.

Berent, Matthew K., Jon A. Krosnick and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys: Do Official Government Records Yield More Accurate Assessments?" *Public Opinion Quarterly* .

Bidwell, Kelly, Katherine Casey and Rachel Glennerster. 2015. "DEBATES: The Impacts of Voter Knowledge Initiatives in Sierra Leone." Working Paper, Stanford Graduate School of Business. URL: `https://www.gsb.stanford.edu/gsb-cmis/gsb-cmis-download-auth/362906`.

Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon and Bin Yu. forthcoming. "Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments." *Proceedings of the National Academy of Sciences* .

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin and Johannes M. Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *The Journal of Human Resources* 32(3):549–576.

Broockman, David and Donald Green. 2014. "Do Online Advertisements Increase Political Can-

didates' Name Recognition or Favorability? Evidence from Randomized Field Experiments." *Political Behavior* 36(2):263–289.

Broockman, David E. and Daniel M. Butler. 2016. "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication." *American Journal of Political Science* .

Broockman, David E. and Joshua L. Kalla. 2016. "Durably reducing transphobia: a field experiment on door-to-door canvassing." *Science* 352(6282):220–224.

Broockman, David, Joshua Kalla and Jasjeet Sekhon. 2017. "Replication Data for: The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs.". doi:10.7910/DVN/EEP5MT, Harvard Dataverse, V1, UNF:6:gM7KTUQ0wCS6voY98ZTw5A==.

Brüggen, E., J. van den Brakel and Jon Krosnick. 2016. "Establishing the accuracy of online panels for survey research." Working paper, available at `https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research`.

Cardy, Emily Arthur. 2005. "An experimental field study of the GOTV and persuasion effects of partisan direct mail and phone calls." *The Annals of the American Academy of Political and Social Science* 601(1):28–40.

Cheung, Paul. 2005. *Designing household survey samples: practical guidelines*. Number 98 *in* "Studies in methods Series F" United Nations.

Collins, Kevin and Joshua Rosmarin. 2016. "Comparing Representativeness in Online and Live Interview Phone Surveys." Presentation at the 2016 meeting of AAPOR.

Conroy-Krutz, Jeffrey and Devra C. Moehler. 2015. "Moderation from Bias: A Field Experiment on Partisan Media in a New Democracy." *Journal of Politics* 77(2):575–587.

Coppock, Alexander. 2016. "Positive, Small, Homogeneous, and Durable: Political Persuasion in Response to Information." Dissertation, Columbia University.

Cubbison, William. 2015. "The Marginal Effects of Direct Mail on Vote Choice." Paper presented at the Annual Meeting of the Midwest Political Science Association. URL: `http://media.wix.com/ugd/3a8c0a_47330c730f56431f8f982a3d842f434a.pdf`.

Dewan, Torun, Macartan Humphreys and Daniel Rubenson. 2014. "The Elements of Political Persuasion: Content, Charisma and Cue." *The Economic Journal* 124(574):F257–F292.

Doherty, David and E. Scott Adler. 2014. "The Persuasive Effects of Partisan Campaign Mailers." *Political Research Quarterly* 67(3):562–573.

Druckman, James N., Donald P. Green, James H. Kuklinski and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(4):627–635.

Druckman, James N. and Thomas J. Leeper. 2012. "Learning More from Political Communication Experiments: Pretreatment and Its Effects." *American Journal of Political Science* 56(4):875–896.

Enos, Ryan D. 2014. "Causal effect of intergroup contact on exclusionary attitudes." *Proceedings of the National Academy of Sciences* 111(10):3699–3704.

Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60(1):123–142.

Fearon, James, Macartan Humphreys and Jeremy M. Weinstein. 2009. "Development Assistance, Institution Building, and Social Cohesion after Civil War: Evidence from a Field Experiment in Liberia." Working Paper.

Funk, Cary and Sara Kehaulani Goo. 2015. A Look at What the Public Knows and Does Not Know About Science. Technical report Pew Research Center.

Gerber, Alan S. 2004. "Does Campaign Spending Work? Field Experiments Provide Evidence and Suggest New Theory." *American Behavioral Scientist* 47(5):541–74.

Gerber, Alan S., Dean Karlan and Daniel Bergan. 2009. "Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions." *American Economic Journal: Applied Economics* 1(2):35–52.

Gerber, Alan S. and Donald P. Green. 2012. *Field experiments: design, analysis, and interpretation*. W. W. Norton.

Gerber, Alan S. and Donald P. Green. 2015. *Get Out the Vote: How to Increase Voter Turnout*. 3 ed. Brookings.

Gerber, Alan S., Gregory A. Huber and Ebonya Washington. 2010. "Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment." *American Political Science Review* 104(4):720–744.

Gerber, Alan S., James Gimpel, Donald Green and Daron Shaw. 2011. "How Large and Long-lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Experiment." *American Political Science Review* 105(1):135–150.

Gooch, Andrew and Lynn Vavreck. 2016. "How Face-to-Face Interviews and Cognitive Skill Affect Item Non-Response: A Randomized Experiment Assigning Mode of Interview." *Political Science Research and Methods* .

Green, Donald P., Alan S. Gerber and David W. Nickerson. 2003. "Getting out the vote in local elections: results from six door-to-door canvassing experiments." *Journal of Politics* 65(4):1083–1096.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22:1–30.

Hall, Thad E. and Betsy Sinclair. 2011. "The American Internet Voter." *Journal of Political Marketing* 10:58–79.

Hartman, Erin, Richard Grieve, Roland Ramsahai and Jasjeet S. Sekhon. 2015. "From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects." *Journal of the Royal Statistical Society, Series A* pp. 1–41.

Heckman, James, Jeffrey Smith and Christopher Taber. 1994. "Accounting for dropouts in evaluations of social experiments." URL: `http://www.nber.org/papers/t0166.pdf`.

Hersh, Eitan D. and Matthew N. Goldenberg. 2016. "Democratic and Republican physicians provide different care on politicized health issues." *Proceedings of the National Academy of Sciences* 113(42):11811–11816.

Hill, Seth J., James Lo, Lynn Vavreck and John R. Zaller. 2007. "The Opt-in Internet Panel: Survey Mode, Sampling Methodology and the Implications for Political Research." Working paper, available at `http://www.allacademic.com/meta/p199541_index.html`.

Himelein, Kristen. 2015. The Socio-Economic Impacts of Ebola in Liberia: Results from a High Frequency Cell Phone Survey, Round 5. Technical report World Bank Group.
URL: *http://www.worldbank.org/content/dam/Worldbank/document/Poverty%20documents/Socio-Economic%20Impacts%20of%20Ebola%20in%20Liberia,%20April%2015%20(final).pdf*

Humphreys, Macartan and Jeremy M. Weinstein. 2012. "Policing Politicians: Citizen Empowerment and Political Accountability in Uganda Preliminary Analysis." Working Paper.

Isbell, Thomas A. 2016. Data Codebook for a Round 6 Afrobarometer Survey in Liberia. Technical report Afrobarometer.
URL: *http://afrobarometer.org/sites/default/files/data/round-6/lib_r6_codebook.pdf*

Iyengar, Shanto and Lynn Vavreck. 2012. Online panels and the future of political communication research. In *The Sage Handbook of Political Communication*. Sage pp. 225–240.

Jackman, Simon and Bradley Spahn. 2015. "Silenced and Ignored: How the turn to voter registration lists excludespeople and opinions from political science andpolitical representation." Working Paper, Stanford University available at `https://www.dropbox.com/s/qvqtz99i4bhdore/silenced.pdf?dl=0`.

Kish, Leslie. 1965. *Survey Sampling*. Wiley.

Kohut, Andrew, Scott Keeter, Carroll Doherty, Michael Dimock and Leah Christian. 2012. "Assessing the Representativeness of Public Opinion Surveys." URL: `http://www.people-press.org/files/legacy-pdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf`.

Lam, Patrick and Kyle Peyton. 2013. "Voter Persuasion in Compulsory Electorates: Evidence from a Field Experiment in Australia." URL: `http://polmeth.wustl.edu/files/polmeth/ausexp.pdf`.

Mann, Christopher B. 2005. "Do Advance Letters Improve Preelection Forecast Accuracy?" *Public Opinion Quarterly* 69(4):561–571.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics* 99(2):210–221.

Michelson, Melissa R. 2016. "The Institutional Review Board: An Approved Project Is Not Always an Ethical Project." *PS: Political Science and Politics* Forthcoming.

Miller, Roy E. and Dorothy L. Robyn. 1975. "A field experimental study of direct mail in a congressional primary campaign: What effects last until election day." *Experimental study of politics* 4(3):1–36.

Nickerson, David W. 2005*a*. "Partisan Mobilization Using Volunteer Phone Banks and Door Hangers." *Annals of the American Academy of Political and Social Science* 601(1):10–27.

Nickerson, David W. 2005*b*. "Scalable protocols offer efficient design for field experiments." *Political Analysis* 13(3):233–252.

Nickerson, David W. 2007. "Don't Talk to Strangers: Experimental Evidence of the Need for Targeting." Presented at the 2007 Annual Meeting of the Midwest Political Scienec Assocation. Available at `https://www.scribd.com/document/98714549/Nickerson-independents`.

Potter, Philip B.K. and Julia Gray. 2008. "Does Costly Signaling Matter? Preliminary Evidence from a Field Experiment." Working paper, available at `http://www.belfercenter.org/sites/default/files/files/publication/Potter%202008%20FINAL%20DOC.pdf`.

Rogers, Todd and David W. Nickerson. 2013. "Can Inaccurate Beliefs about Incumbents Be Changed? And Can Reframing Change Votes?" Working Paper RWP13-018,

Harvard Kennedy School. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2271654.

Sadin, Meredith L. 2014. Campaigning with Class: Ambivalent Stereotypes and Candidate Wealth in U.S. Elections PhD thesis Princeton.

Sävje, Fredrik, Michael Higgins and Jasjeet S. Sekhon. 2016. "Improving Massive Experiments with Threshold Blocking." *Proceedings of the National Academy of Sciences* 113(27):7369–7376.

Shineman, Victoria Anne. 2016. "If You Mobilize Them, They Will Become Informed: Experimental Evidence that Information Acquisition Is Endogenous to Costs and Incentives to Participate." *British Journal of Political Science* .

Sniderman, Paul M. and Douglas B. Grob. 1996. "Innovations in Experimental Design in Attitude Surveys." *Annual Review of Sociology* 22:377–399.

Solomon, Richard L. 1949. "An extension of the control group design." *Psychological bulletin* 46(2):137–150.

Strauss, Aaron B. 2009. Political Ground Truth: How Personal Issue Experience Counters Partisan Biases PhD thesis Princeton.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Appendix

Figure A1: Example Experimental Design Using All Four Practices